DOCUMENT RESUME

ED 346 117 TM 018 360

AUTHOR Carlson, James E.; Jirele, Tom
TITLE Dimensionality of 1990 NAEP Mathematics Data.
PUB DATE Apr 92
NOTE 63p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Comparative Testing; *Computer Software Evaluation; *Factor Analysis; Grade 4; Grade 8; Intermediate Grades; *Item Response Theory; Junior High Schools; *Mathematics Tests; National Surveys; Research Methodology; Scaling
IDENTIFIERS *Dimensionality (Tests); *National Assessment of Educational Progress; NOHARM Computer Program; TESTFACT

ABSTRACT

Some results are presented relating to the dimensionality of the 1990 National Assessment of Educational Progress (NAEP) mathematics item-response data. Based on theoretical considerations, practical limitations, and previous research, two procedures were selected for study: full information factor analysis as implemented in the TESTFACT computer program and normal harmonic factor analysis as implemented in the NOHARM computer program. One-dimensional, two-dimensional, and three-dimensional models were fitted using both programs. NOHARM analysis used 941 fourth-grade examinees with 50 items, 914 fourth-grade examinees with 42 items, 917 eighth-grade examinees with 54 items, and 924 eighth-grade examinees with 58 items. Sixteen TESTFACT analyses were conducted with real data sets. Comparisons show better fit when data are known to be one dimensional than in the case of real NAEP data. The results seem consistent over both programs and over different data sets. There is, however, evidence of a strong first dimension fitting the examinee-item interaction data. Overall, it is suggested that data are sufficiently unidimensional to support use of a composite scale for describing NAEP data, although there is evidence that two dimensions would better fit the data than would one dimension. Eight tables and 23 graphs present analysis results, and there is a 33-item list of references. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
JAMES E. CARLSON
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED346117

Dimensionality of 1990 NAEP Mathematics Data

James E. Carlson

Tom Jirele

Educational Testing Service

Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April, 1992.

TM018360

BEST COPY AVAILABLE

Dimensionality of 1990 NAEP Mathematics Data[1]

James E. Carlson

Tom Jirele

Educational Testing Service

Introduction

The dimensionality of dichotomously-scored test items has been studied by many researchers using a variety of techniques over the years. McDonald (1981, 1982a, 1982b, 1985) reviewed much of the theoretical literature in this area; Hattie (1984, 1985) reviewed and empirically studied various indices for determining unidimensionality. Muthén (1991), Rock (1991), and Zwick (1986, 1987) studied the dimensionality of portions of the NAEP data using different techniques. In this paper some results are presented relating to dimensionality of the 1990 NAEP mathematics item-response data.

Theoretical or empirical studies of dimensionality that involve statistical/psychometric techniques involve item-response data resulting from the examinee-item interaction and **not** the dimensionality of items as entities separate from examinees. Hence in this paper we will refer to the dimensionality of a set of item response data, with the understanding that such data result from the examinee-item interaction in a specific population.

Methods of Assessing Dimensionality

McDonald (1962, 1967, 1981, 1982a, 1982b, 1985) makes a strong case for the use of nonlinear factor analysis and the examination of residuals in order to assess the dimensionality of item-response data. His case, which is based on theoretical considerations has been supported by empirical studies, both in his own research and that of Hattie (1984, 1985) and Knol & Berger (1991).

Work by Bock and his colleagues (Bock & Aitkin, 1981; Bock, Gibbons, & Muraki, 1988; Bock & Lieberman, 1970), and by Christoffersson (1975), and Muthén (1978) suggests that factor analysis procedures specifically designed for dichotomous data are useful in assessing dimensionality. McDonald (1982b) showed that Christoffersson's method is directly related to his own nonlinear method of factor analysis. Mislevy (1986) also discussed these techniques and Knol and Berger (1991) found that they worked well with simulated data. Zwick (1986, 1987) studied these techniques both with NAEP data and with simulated data.

Stout (1983, 1987, 1990) and Rosenbaum (1984, 1985, see also Holland & Rosenbaum, 1986) have both developed techniques for formally testing hypotheses relating to dimensionality. Stout's procedures involve fitting factor analytic models and performing tests about the dimensionality, while

---

[1]Thanks are due to many colleagues for their suggestions, etc. In particular the authors would like to thank Bob Mislevy, Eiji Muraki and Ming-mei Wang for their help in clarifying several issues, Nancy Allen, John Donoghue, Don Rock, and Rebecca Zwick for suggestions about specific procedures.

Rosenbaum's technique provides a test of local independence. Little work seems to have been done in applying Stout's method, but Zwick (1986, 1987) included Rosenbaum's procedures in her study.

Previous analyses of NAEP data

Zwick (1986, 1987) studied NAEP reading data collected during the 1983-84 academic year. She also examined simulated data designed to mirror the NAEP reading item-response data but having known dimensionality. Principal components analysis (PCA) was applied to both phi and tetrachoric correlation matrices. Full information item factor analysis (Bock & Aitkin, 1981; Bock, Gibbons, & Muraki, 1988) implemented in the TESTFACT computer program (Wilson, Wood, & Gibbons, 1991) was applied to portions of the dataset, as was Rosenbaum's (1984, 1985) tests. The latter two procedures were applied to portions of the dataset because of the substantial computer resources required to do such analyses on the full NAEP datasets. One reason for analyzing the simulated datasets was to determine whether the BIB Spiraling design had any effect on the dimensionality analyses. Zwick concluded that there was substantial agreement between the PCA and the full information factor analysis and Rosenbaum results, and that the results using BIB spiraling were similar to results for complete datasets. Her overall conclusion was that "it is not unreasonable to treat the data as unidimensional (1987, p. 306)."

Rock (1991) investigated "whether the presently reported subscale scores do span a multidimensional space defined by the content area subscales at each of the three grade levels in mathematics and science (p. 1)." In order to study the discriminant validity he formed two parcels of items that are homogeneous with respect to content for each subtest of the NAEP mathematics and science tests from the 1990 assessment, and studied their dimensionality using confirmatory factor analysis. Factor intercorrelations averaged across booklets ranged from .86 to .95 in mathematics, and from .94 to .96 in science. Rock concluded that there was little evidence for discriminant validity except for the geometry subscale at the 8th grade level, and that "we are doing little damage in using a composite score in mathematics and science (p. 2)."

Muthén (1991) further analyzed Rock's mathematics data using a second-order factor model and also examined subgroup differences in dimensionality. Although he found evidence of content-specific variation within subgroups, the average (across 7 booklets) percentages of such variation ranged from essentially zero to 22, and two-thirds of these percentages were smaller than 10.

Methods

The nature of the NAEP datasets limits the applicability of some computer programs that can be used for assessing dimensionality. The resources required to perform analyses by some programs make them prohibitively expensive for extensive use. Additional problems for the assessment of dimensionality arise as a result of the design for data collection and the fact that items that are omitted by examinees at the end of test booklets are considered as not administered during the scaling procedures. Many of the computer programs cannot readily deal with the "missing data" resulting from this design. Other programs, for which source code was available, have been modified, and some can analyze product-moment matrices computed from missing data.

NAEP test booklets are comprised of blocks of items. These blocks are paired and administered using a balanced incomplete block (BIB) spiraling design (Beaton, Johnson, & Ferris, 1987; Zwick, 1987). Hence no examinee is administered a complete set of all items in a subject area (or in any subscale of a

subject area). The design, which is efficient for purposes of estimating group mean proficiency, precludes performing dimensionality analyses of the entire set of items. The incomplete nature of the entire dataset, with blocks of data missing by design, would result in separate dimensions being identified within each block by any of the techniques used in this study. In mathematics the subscale scores are combined into composites for purposes of reporting results. Hence the question of whether there is a strong first dimension underlying the item-response data is important.

Based on theoretical considerations, practical limitations, and previous research, two procedures were selected for this study, full information item factor analysis (Bock, Gibbons, & Muraki, 1988) as implemented in the TESTFACT computer program (Wilson, Wood, & Gibbons, 1991), and normal harmonic factor analysis (McDonald, 1962, 1967, 1981) as implemented in the NOHARM program (Fraser, 1988). Because they might be of interest to some readers, latent roots of the various matrices are also reported.

The data analyzed were the item-response data from selected 1990 NAEP main assessment mathematics tests. Item response data from four booklets from the BIB spiral design in mathematics (two each at grades 4 and 8) were studied. In addition four simulated one-dimensional datasets were analyzed in order to compare analyses of actual NAEP data with similar data of known dimensionality.

The NOHARM and TESTFACT programs require as input the fixed values of the c parameter in the three-parameter IRT model underlying the procedures. Two different approaches were used for specifying these parameters. In the first approach the parameter estimates derived by using the BILOG (Mislevy & Bock, 1982) computer program to scale the items were specified. These are the c parameters reported in the *NAEP 1990 Technical Report* (Johnson & Allen, 1992), and result from separately scaling each of the five subtests. In the second approach all c parameters were arbitrarily set to 0.1. Additionally, two different approaches were tried for items that examinees omitted at the end of the booklets. First such items were scored as "incorrect" so that there was complete data on each examinee for each item. As a second technique, following usual NAEP practice, such items were taken to be "not reached" and treated as missing data.

NOHARM Analyses: One-, two-, and three-dimensional models were fitted using the NOHARM program. Exploratory factor analysis models were specified. As recommended by McDonald the fit of the model to the data was studied by examination of residuals.

The NOHARM program has a default number of iterations after which, if convergence has not been reached, it stops and prints out the solution. The default convergence criterion of a change of .00001 in the fitting function, which is rather rigorous, was employed.

TESTFACT Analyses: One-, two-, and three-dimensional models were fitted using the TESTFACT program. Like those fitted with NOHARM the TESTFACT models are exploratory factor analytic models. Sizes of the chi-square difference statistics (between models of different dimensions) and the AIC index (Akaike, 1987) were used to assess dimensionality.

TESTFACT uses numerical quadrature in fitting models to the data. In all cases the default numbers of quadrature points was used. These are 10 points for one factor, five points per dimension (resulting in $5^2 = 25$ points) for two factors, and three points per dimension for three factors.

In the EM algorithm employed by TESTFACT the defaults of five E-step iterations and 5 M-step iterations were used. The default of .005 for a precision criterion for convergence in the M-steps was also used. No prior distributions were specified on the slope and intercept parameters.

Latent roots: Latent roots were computed from certain matrices related to the NOHARM and TESTFACT procedures.

Latent roots were computed from the uncentered product-moment matrix of each data set. The latter is the matrix that is input to NOHARM. It should be noted, however, that the roots involve fitting linear functions to the matrix whereas NOHARM fits nonlinear functions. Thus the root plots for this matrix should not be construed as relating to McDonald's techniques.

Latent roots of tetrachoric correlation matrices produced by TESTFACT are also reported. The correlation coefficients in these matrices are corrected for small frequencies (Wilson, Wood, & Gibbons, 1991). Divgi's (1979) method is used to compute the tetrachorics, and, using the supplied (fixed) values of the c parameters, Carroll's (1945) correction for guessing is applied. It is the roots of these corrected correlations that were examined in this study. TESTFACT uses these coefficients only for computing starting values for the full information item factor analysis iterations. Hence the roots are not directly related to dimensionality as assessed in this study with TESTFACT.

Results

NOHARM Analyses: The residual results from the runs with NOHARM are displayed in Tables 1a through 1d and Figures 1a through 1d for two booklets of Mathematics items at the fourth grade level, and two at the eighth grade level. The fourth-grade datasets resulted from the interactions of 941 examinees with 50 items, and 914 examinees with 42 items, respectively. The eighth-grade datasets resulted from the interactions of 917 examinees with 54 items and 929 examinees with 58 items, respectively. There was a strong tendency for the three-dimensional solutions not to converge within the default number of iterations.

As may be seen from Tables 1a through 1d and Figures 1a through 1d the fit to the data, as judged from the root-mean-square residuals tends to differ a little depending on the fixed values of c parameters. Recall that they were fixed to a common value of .1 for some runs and to the BILOG c parameters for others. For Grade 4, Booklet 11, the fit was a little better with .1 for NAEP data and a little better with BILOG values for the simulated data. For the other three datasets the fit was better with the BILOG values for complete and simulated data, and a little better using .1 for c parameters when not-reached responses were treated with the pairwise deletion method. As might be expected the fit for the simulated data was better than that for the real data. There is also a tendency for the increase in fit to be at about the same rate between fitting one and two factors as between fitting two and three factors for the simulated data. For the NAEP data, on the other hand, there is a slight tendency for a larger increase in fit between fitting one and two factors than between fitting two and three factors. In the figures "BlgC" refers to the analyses specifying BILOG values for c parameters, ".1C" refers to specifying those parameters as .1, "Pwd" refers to the use of the pairwise deletion technique, and "Sim" refers to the simulated data. In interpreting the figures note that the scale differs slightly from figure to figure and the values all lie between .0045 and .01, a small range.

Table 1a

NOHARM Root Mean Square Residuals: Mathematics Grade 4 Booklet 11 (941 Examinees, 50 Items)

Dataset	Number of Factors	BILOG c parameters	.1 c parameters
NAEP Data: complete	1	.00871	.00720
	2	.00813	.00633
	3	.00787[a]	.00581[a]
NAEP Data: pairwise deletion	1	.00773	.00731
	2	.00706	.00644
	3	.00676[a]	.00587[a]
Simulated Data	1	.00606	.00637
	2	.00568	.00585
	3	.00543[a]	.00552[a]

[a] did not converge

Table 1b

NOHARM Root Mean Square Residuals: Mathematics Grade 4 Booklet 12 (914 Examinees, 42 Items)

Dataset	Number of Factors	BILOG c parameters	.1 c parameters
NAEP Data: complete	1	.00654	.00677
	2	.00598	.00625
	3	.00556	.00583
NAEP Data: pairwise deletion	1	.00753	.00694
	2	.00688	.00626
	3	.00662[a]	.00599[a]
Simulated Data	1	.00656	.00689
	2	.00619	.00653
	3	.00583	.00619

[a] did not converge

Table 1c

NOHARM Root Mean Square Residuals: Mathematics Grade 8 Booklet 10 (917 Examinees, 54 Items)

Dataset	Number of Factors	BILOG c parameters	.1 c parameters
NAEP Data: complete	1	.00808	.00851
	2	.00714[a]	.00761
	3	.00643[a]	.00696[a]
NAEP Data: pairwise deletion	1	.00898	.00850
	2	.00841	.00791
	3	.00803[a]	.00758[a]
Simulated Data	1	.00544	.00556
	2	.00500	.00508
	3	.00477[a]	.00486[a]

[a] did not converge

Table 1d

NOHARM Root Mean Square Residuals: Mathematics Grade 8 Booklet 13 (929 Examinees, 58 Items)

Dataset	Number of Factors	BILOG c parameters	.1 c parameters
NAEP Data: complete	1	.00761	.00677
	2	.00649	.00625
	3	.00622[a]	.00583
NAEP Data: pairwise deletion	1	.01000	.00694
	2	.00928	.00626
	3	.00896[a]	.00599[a]
Simulated Data	1	.00566	.00681
	2	.00529	.00643
	3	.00507[a]	.00626[a]

[a] did not converge

Figure 1a

NOHARM RMS Residuals Grade 4 Book 11

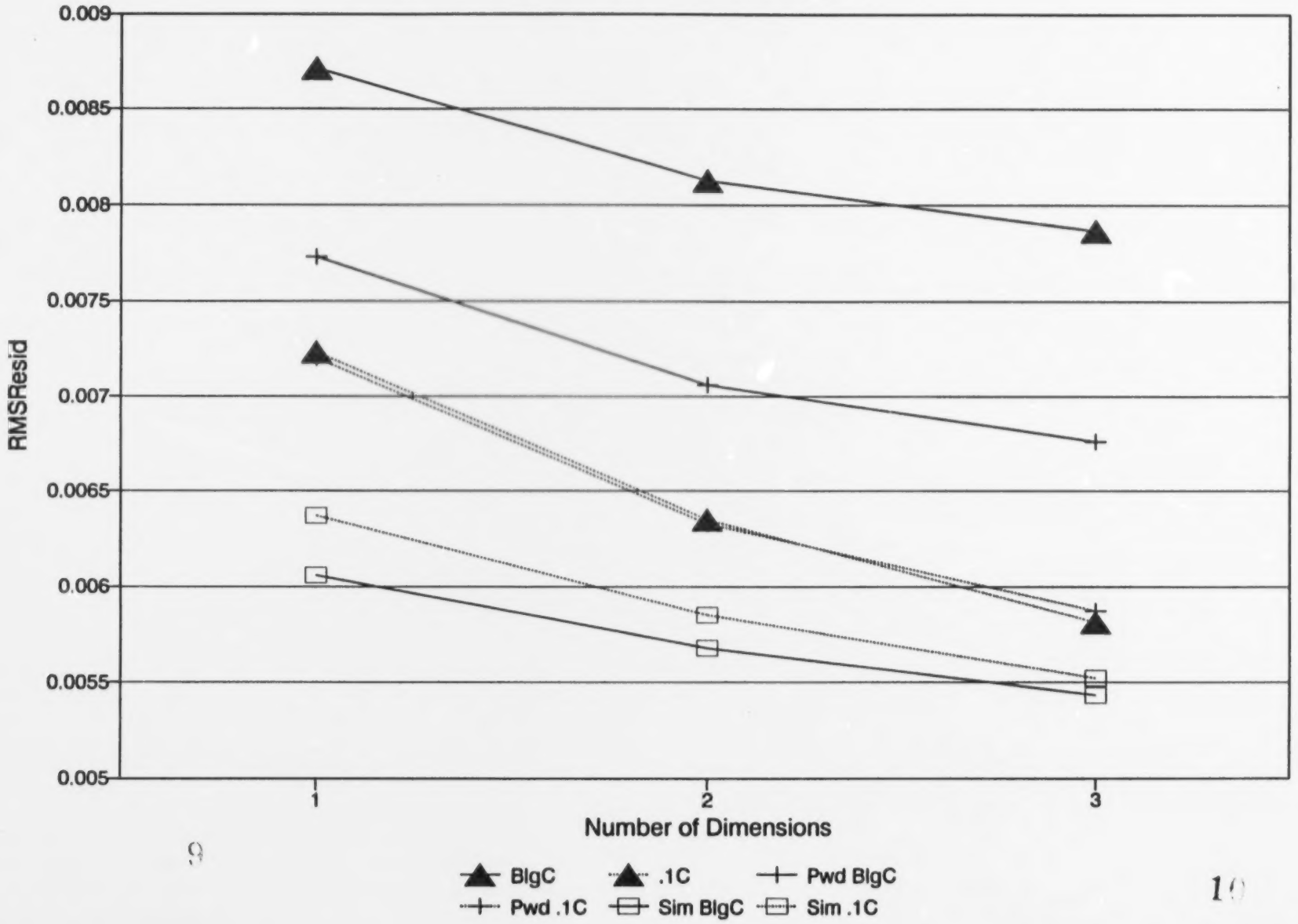


Figure 1b

NOHARM RMS Residuals Grade 4 Book 12

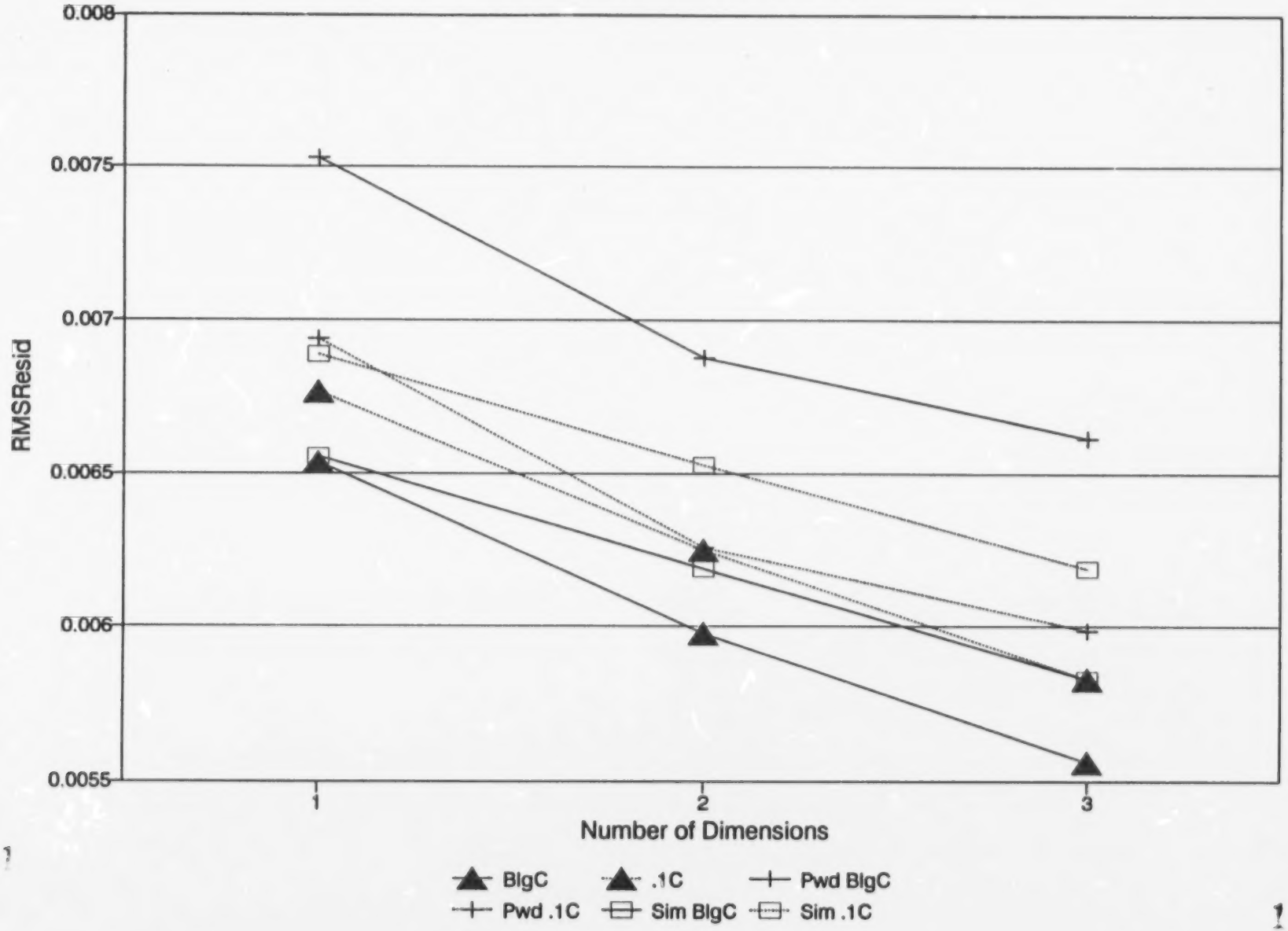


Figure 1c

NOHARM RMS Residuals Grade 8 Book 10

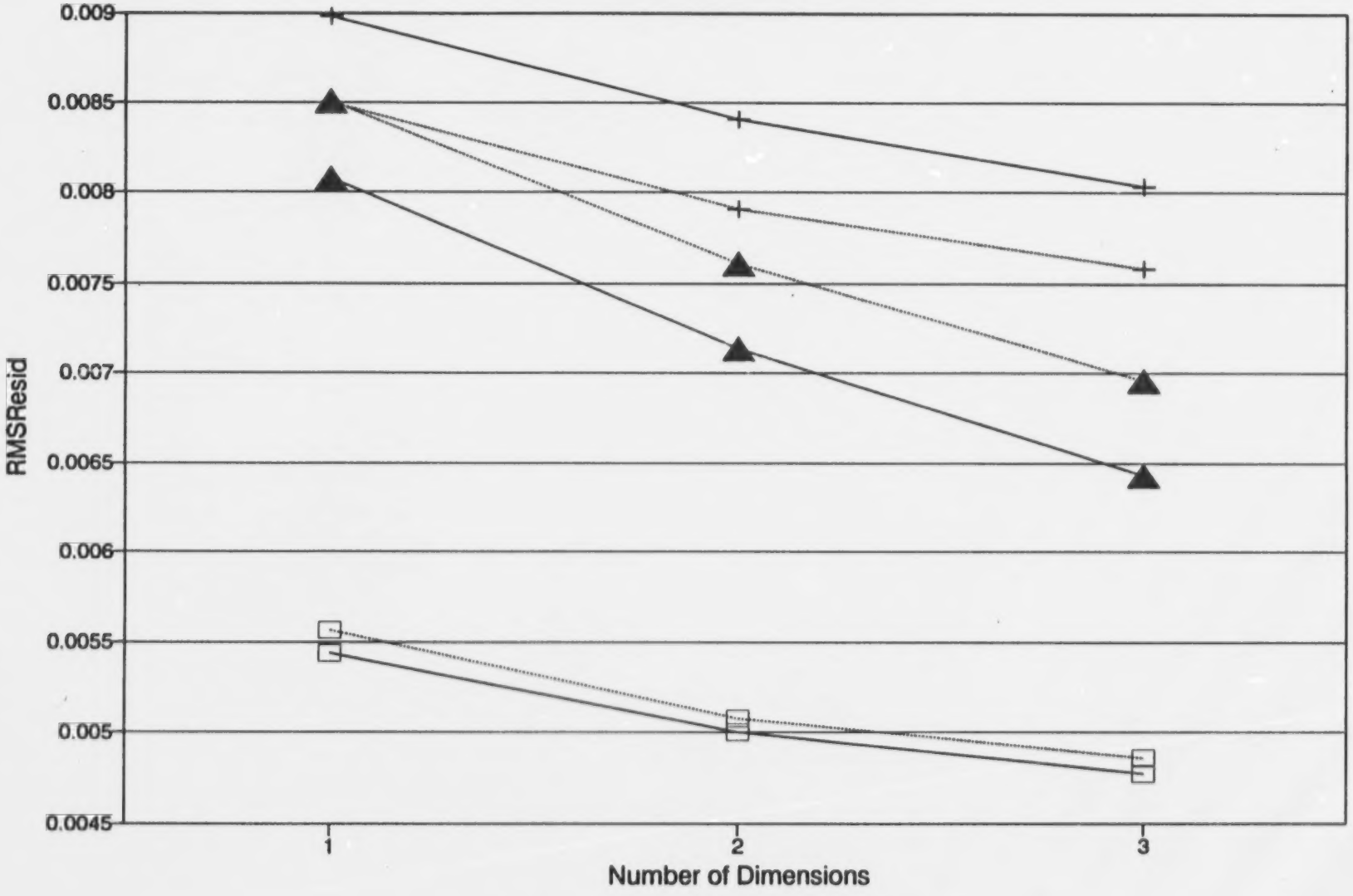
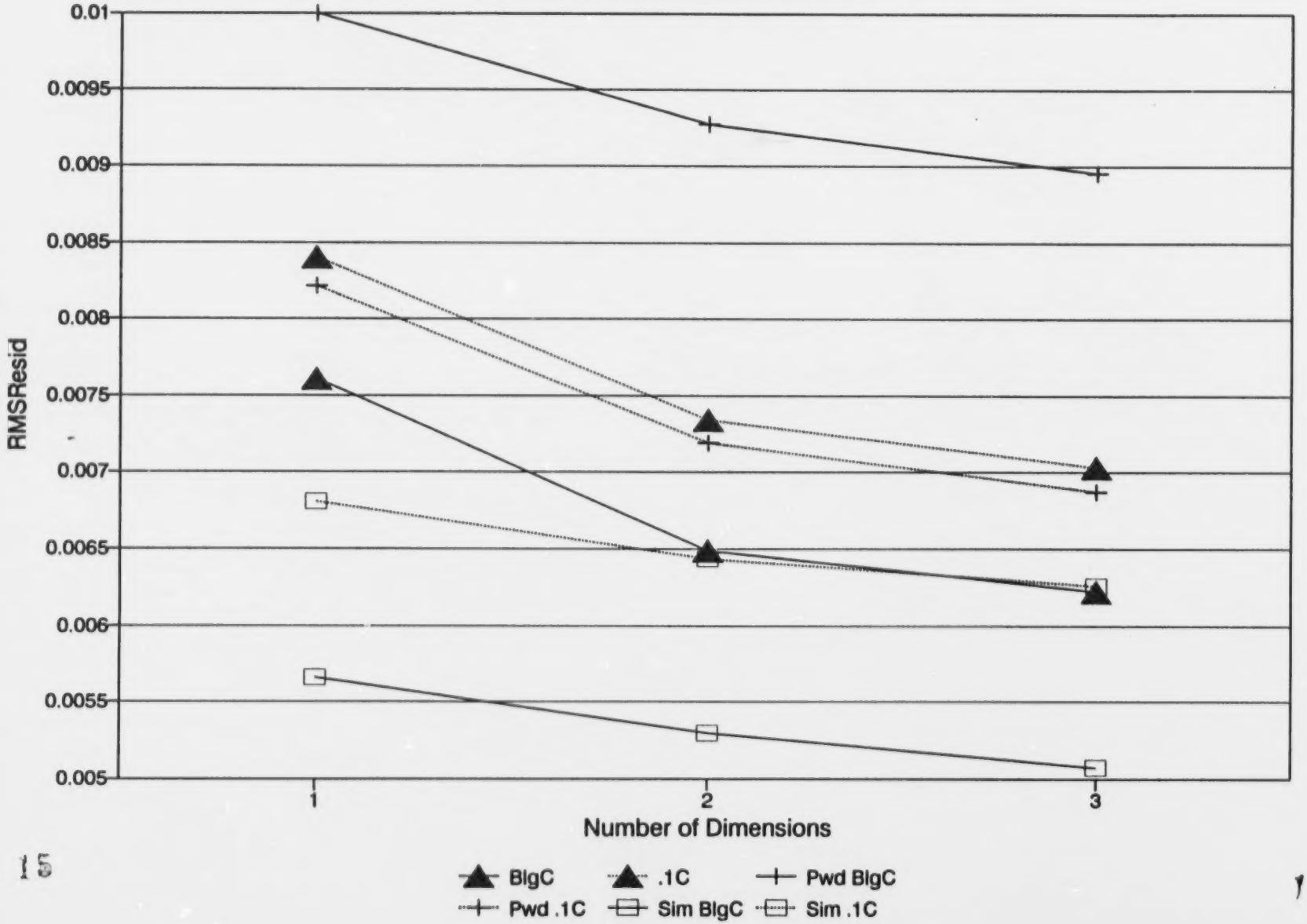


Figure 1d

NOHARM RMS Residuals Grade 8 Book 13



TESTFACT Analyses: The TESTFACT results are presented in Tables 2a through 2d. It should be noted in comparing the statistics reported in these tables that differences in numbers of degrees of freedom make some comparisons difficult. The reason for the differences is that the number of degrees of freedom depends on the actual cell frequencies in cross tabulations over items. The AIC indices are shown in Figures 2a through 2d. In five of the sixteen TESTFACT analyses, as pointed out in the tables, it was necessary to specify a beta prior distribution on the item parameters in order to get solutions for some numbers of dimensions. In the other 11 analyses no prior distribution was specified. In cases where there were solutions with and without priors there was very little difference (less than 1.0%) in the statistics reported in this paper.

Out of 16 TESTFACT analyses that were conducted on real NAEP datasets the likelihood criterion failed to increase between fitting one and two factors two times and between fitting two and three factors eleven times. These results strongly suggest no more than two dimensions underlying any of the datasets. For the simulated data these results are more pronounced than for the real data, with six out of eight analyses resulting in no increase between one and two factors. The other two had increases all the way to three factors.

When using the AIC index to assess fit of a model (Akaike, 1987) the smallest value of the index represents the best fit. Using this criterion for the real data one dimension would be concluded to underlie the data in ten analyses and two factors in the remaining six analyses. In all eight analyses of simulated data the AIC was a minimum for one factor so this index performed very well.

Using the differences in the chi square statistics between fitting one and two factors for real data, 12 analyses resulted in a significant difference at the .05 level. As pointed out above the value of the likelihood function failed to increase between models of one and two factors in two other analyses. In no analyses was there a significant decrease in chi square between fitting two and three factors in the real data. These results strongly suggest only one or two dimensions underlying the datasets. For the simulated data the results were stronger with only three significant differences, which occurred between two and three dimensional models, and these only occurred after the likelihood function value had failed to increase between one and two factors. Thus these statistics performed very well for detecting the true one-dimensional nature of the simulated data.

Table 2a

TESTFACT Results: Grade 4, Booklet 11 (941 Examinees, 50 Items)

Dataset	No. Factors	Chi Square	Degrees of Freedom	AIC
NAEP Blg C	1	37,387.57	839	35,709.57
	2[a,b]	37,029.97	790	35,449.97
	3[c]	37,277.35	742	35,793.35
NAEP .1 C	1[a]	37,131.00	839	35,453.00
	2[b]	37,050.93	790	35,470.93
	3[c]	37,515.32	742	36,031.32
NAEP[d] Blg C Pw Del	1[a]	36,429.73	839	34,751.73
	2[b]	36,358.27	790	34,778.27
	3[c]	36,516.10	742	35,032.10
NAEP .1 C Pw Del	1[a]	36,512.04	839	34,834.04
	2[b]	36,437.42	790	34,857.42
	3[c]	36,715.51	742	35,231.51
Simulated Data Blg C	1[a]	37,166.85	776	35,614.85
	2[c]	37,264.46	727	35,810.46
	3[b]	37,128.66	679	35,770.66
Simulated Data .1 C	1[a]	37,382.05	776	35,830.05
	2[c]	37,431.20	727	35,977.20
	3[c]	37,605.16	679	36,247.16

[a] best fitting by AIC criterion
[b] significant change in chi square from 1 fewer factors at the .05 level
[c] no increase in likelihood function over 1 fewer factors
[d] beta prior on item parameters necessary to get solution

Table 2b

TESTFACT Results: Grade 4, Booklet 12 (914 Examinees, 42 Items)

Dataset	No. Factors	Chi Square	Degrees of Freedom	AIC
NAEP Blg C	1[a]	29,142.67	828	27,486.67
	2[b]	29,073.47	787	27,499.47
	3[c]	29,593.25	747	28,099.25
NAEP .1 C	1[a]	29,197.73	828	27,541.73
	2[c]	29,259.65	787	27,685.65
	3[c]	29,270.97	747	27,776.97
NAEP[d] Blg C Pw Del	1[a]	28,642.22	828	26,986.22
	2	28,611.88	787	27,037.88
	3[c]	28,633.58	747	27,139.58
NAEP[e] .1 C Pw Del	1[a]	28,735.89	828	27,079.89
	2[c]	28,782.85	787	27,208.85
	3	28,777.09	747	27,283.09
Simulated Data Blg C	1[a]	29,691.09	744	28,203.09
	2[c]	29,727.91	703	28,321.91
	3[c]	29,760.99	663	28,434.99
Simulated Data .1 C	1[a]	29,869.73	744	28,381.73
	2[c]	29,898.06	703	28,492.06
	3[c]	29,959.89	663	28,633.89

[a] best fitting by AIC criterion
[b] significant change in chi square from 1 fewer factors at the .05 level
[c] no increase in likelihood function over 1 fewer factors
[d] beta prior on item parameters necessary for two-dimensional solution
[e] beta prior on item parameters necessary for three-dimensional solution

Table 2c

TESTFACT Results: Grade 8, Booklet 10 (917 Examinees, 54 Items)

Dataset	No. Factors	Chi Square	Degrees of Freedom	AIC
NAEP[d] Blg C	1	36,483.56	808	34,867.56
	2[a,b]	36,302.53	755	34,792.53
	3	36,263.94	703	34,857.94
NAEP .1 C	1	36,588.53	808	34,972.53
	2[a,b]	36,444.53	755	34,934.53
	3[c]	36,575.88	703	35,169.88
NAEP Blg C Pw Del	1[a]	35,052.77	808	33,436.77
	2[b]	34,973.32	755	33,463.32
	3[c]	35,122.03	703	33,716.03
NAEP .1 C Pw Del	1[a]	35,182.88	808	33,566.88
	2	35,112.73	755	33,602.73
	3	35,061.19	703	33,655.19
Simulated Data Blg C	1[a]	35,602.63	763	34,076.63
	2[c]	35,666.29	710	34,246.29
	3	35,643.62	658	34,327.62
Simulated Data .1 C	1[a]	35,743.21	763	34,217.21
	2	35,695.17	710	34,275.17
	3	35,647.92	658	34,331.92

[a] best fitting by AIC criterion
[b] significant change in chi square from 1 fewer factors at the .05 level
[c] no increase in likelihood function over 1 fewer factors
[d] beta prior on item parameters necessary for three-dimensional solution

Table 2d

TESTFACT Results: Grade 8, Booklet 13 (929 Examinees, 58 Items)

Dataset	No. Factors	Chi Square	Degrees of Freedom	AIC
NAEP Blg C	1	41,811.77	811	40,189.77
	2[a,b]	41,649.13	754	40,141.13
	3[c]	42,592.87	698	41,196.87
NAEP .1 C	1	41,986.79	811	40,364.79
	2[a,b]	41,641.67	754	40,133.67
	3[c]	41,919.84	698	40,523.84
NAEP Blg C Pw Del	1[a]	41,383.81	811	39,761.81
	2[b]	41,297.60	754	39,789.60
	3	41,286.01	698	39,890.01
NAEP[d] .1 C Pw Del	1	41,454.45	811	39,832.45
	2[a,b]	41,226.39	754	39,718.39
	3	41,154.33	698	39,758.33
Simulated Data Blg C	1[a]	41,721.75	764	40,193.75
	2[c]	41,935.20	707	40,521.20
	3[b]	41,792.01	651	40,490.01
Simulated Data .1 C	1[a]	41,963.21	764	40,435.21
	2[c]	42,137.20	707	40,723.20
	3[b]	42,013.78	651	40,711.78

[a] best fitting by AIC criterion
[b] significant change in chi square from 1 fewer factors at the .05 level
[c] no increase in likelihood function over 1 fewer factors
[d] beta prior on item parameters necessary for three-dimensional solution

Figure 2a

Grade 4 Book 11 AIC Index

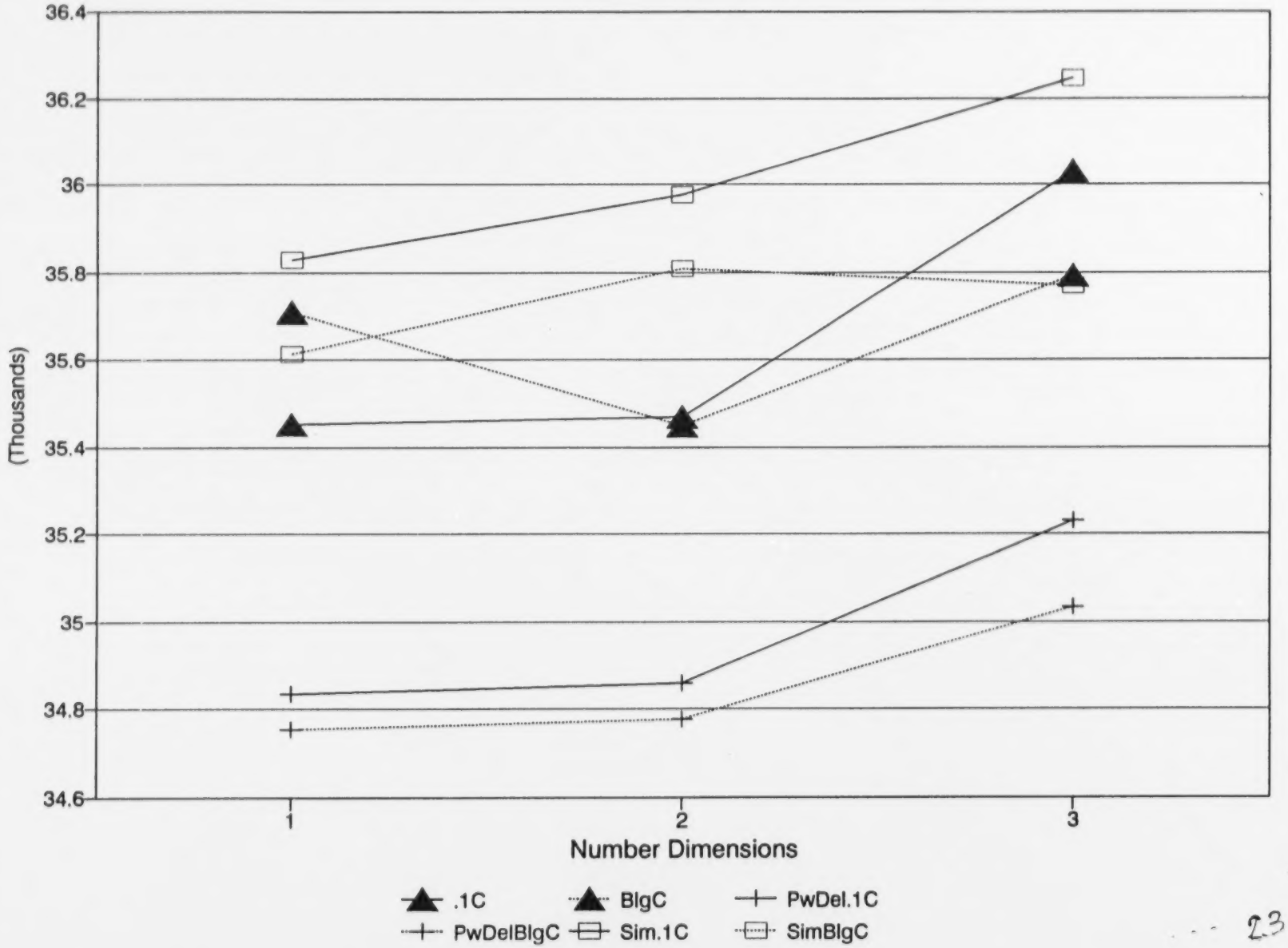


Figure 2b

Grade 4 Book 12 AIC Index

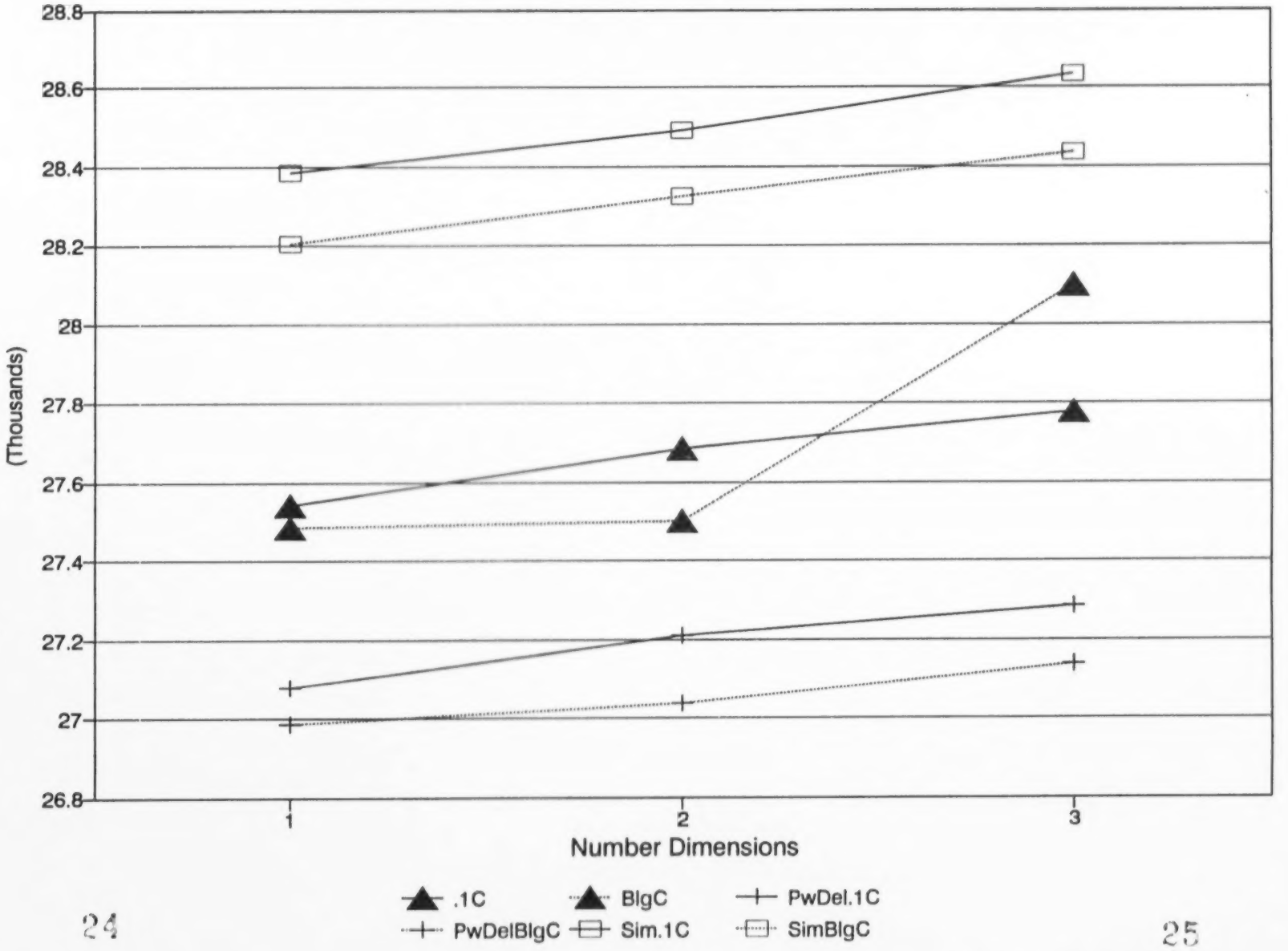


Figure 2c

Grade 8 Book 10 AIC Index

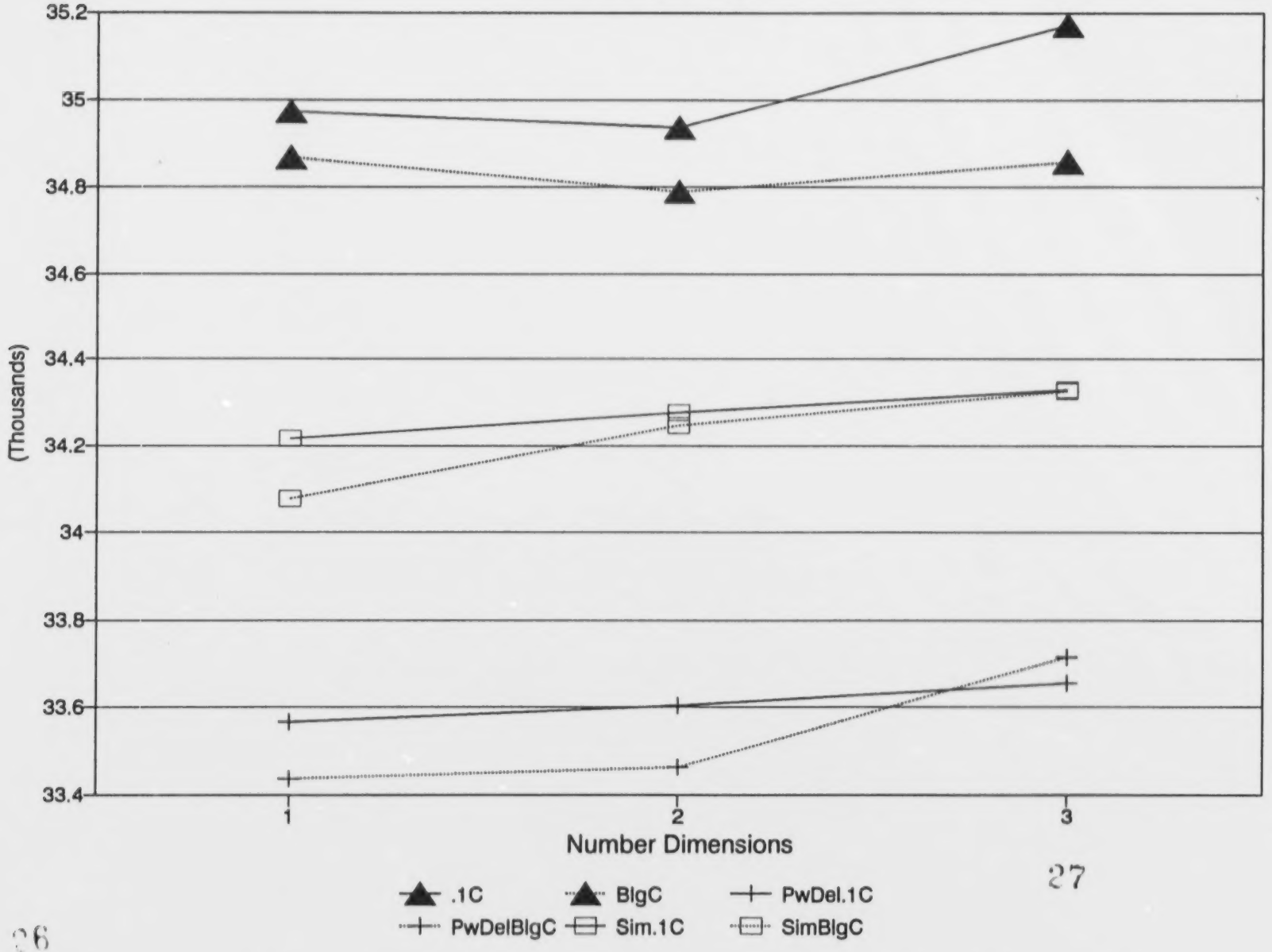
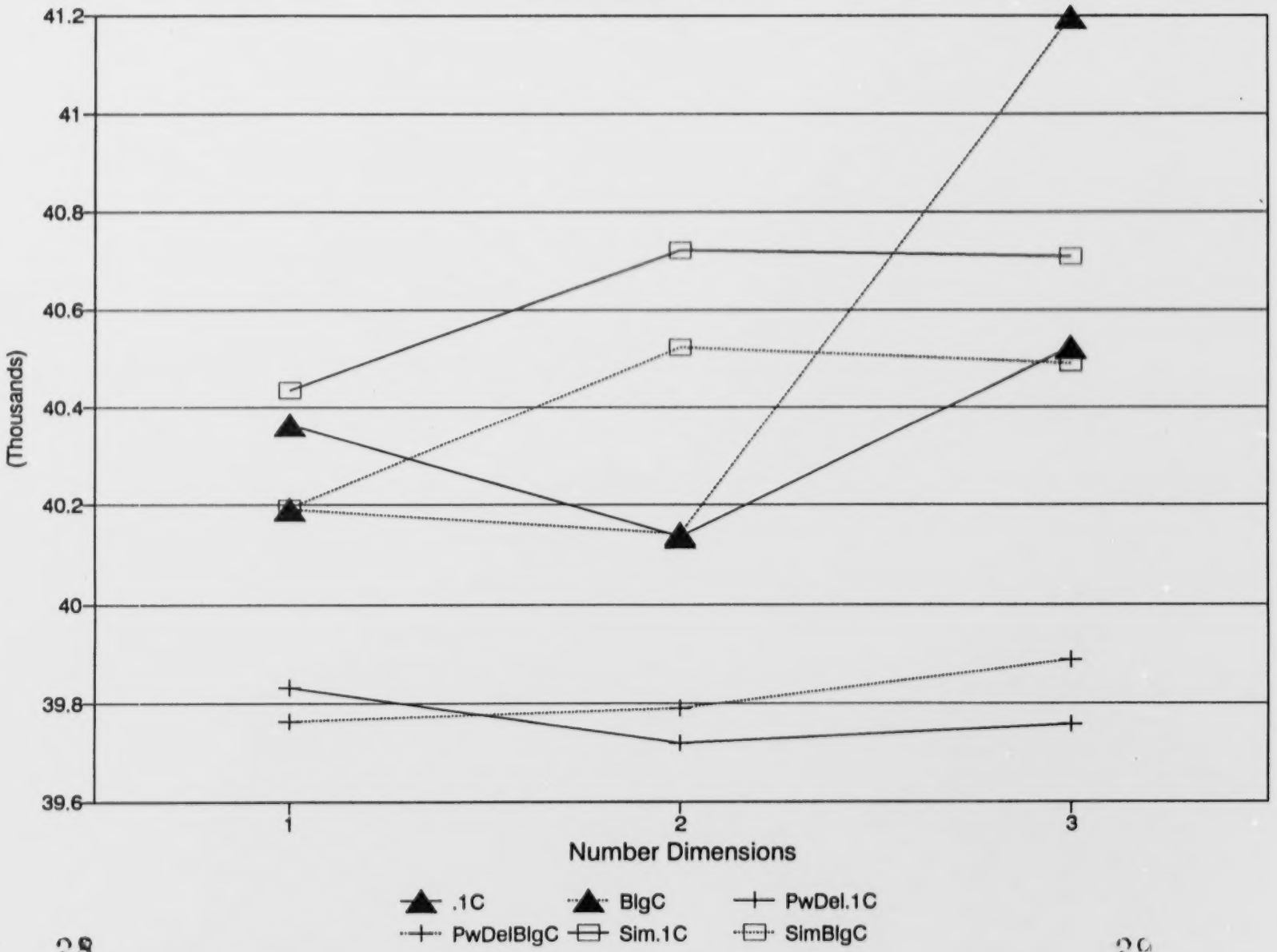


Figure 2d

Grade 8 Book 13 AIC Index



Latent Root Plots

Latent root plots for the first ten roots of the various matrices of interest are displayed in Figures 3a through 3e. Because the scales of the plots make it difficult to differentiate among the various curves a second set of plots, Figures 4a through 4e, are used to display the 2nd through 10th roots of each matrix. In these Figures "Pwd" refers to the tetrachoric correlation matrix formed using pairwise deletion of not-reached item responses, "Sim" to the simulated data, "BlgC" to use of BILOG c parameters as the "guessing" parameter, ".1C' to the use of .1 as the "guessing" parameter, and "SSCP/N" refers to the uncentered product-moment matrix. In the SSCP/N plots the grade and booklet are identified, for example, Grade 4 Booklet 11 by "G4B11".

The latent root structure of the 8th grade booklet 13 tetrachoric matrix using BILOG c parameters and pairwise deletion (Figures 3d, 4d) differs markedly from that of the other matrices. It has a very large first root and relatively small subsequent roots. Also, the tetrachoric matrices resulting from the use of BILOG c parameters for Grade 4 Booklet 12, with and without pairwise deletion, have 3 roots that are somewhat larger than those of other tetrachoric matrices (Figures 3b, 4b).

Figure 3a

Grade 4 Book 11 First Ten Latent Roots

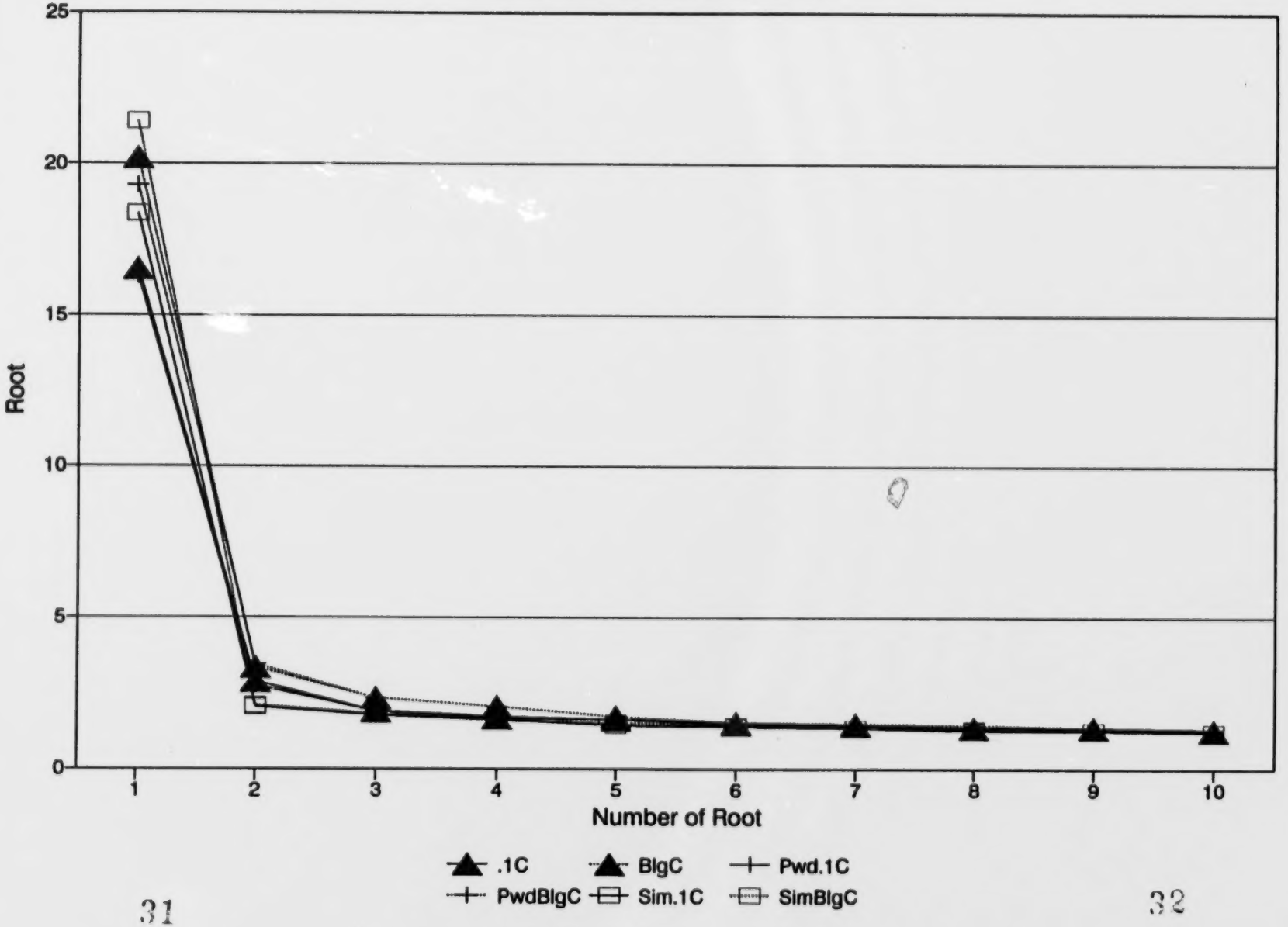


Figure 3b

Grade 4 Book 12 First Ten Latent Roots

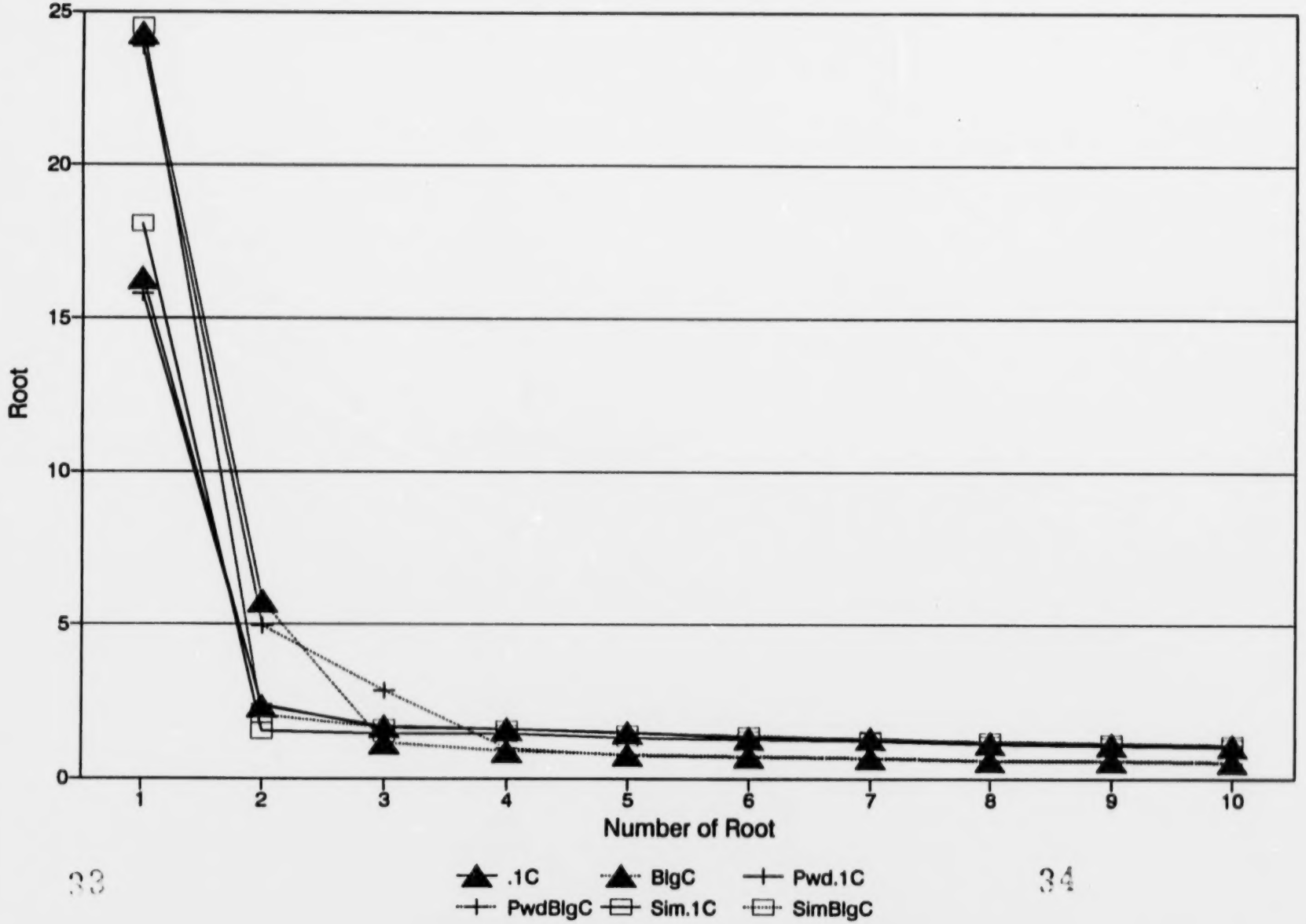


Figure 3c

Grade 8 Book 10 First Ten Latent Roots

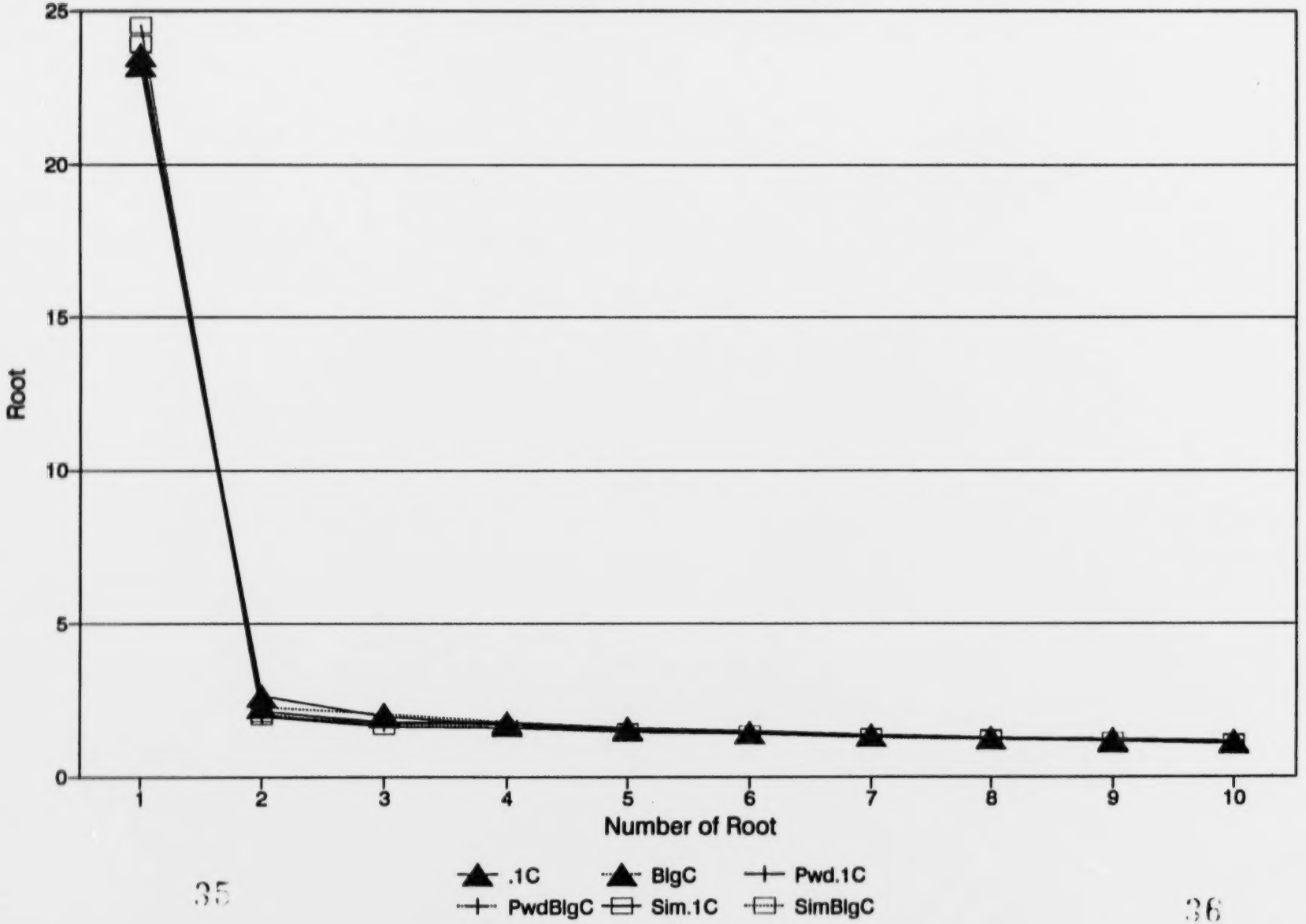


Figure 3d

Grade 8 Book 13 First Ten Latent Roots

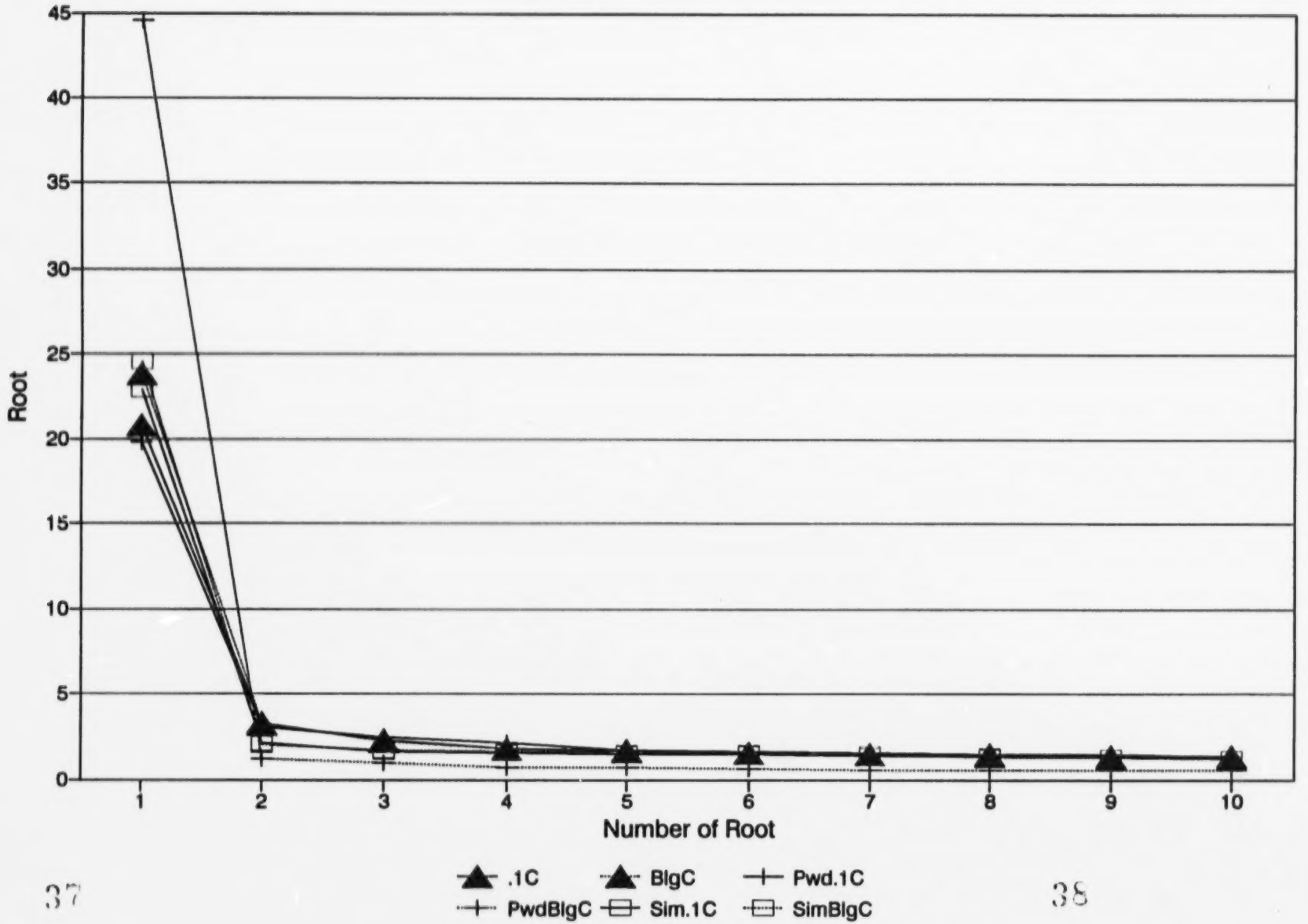


Figure 3e

First Ten Roots of SSCP/N

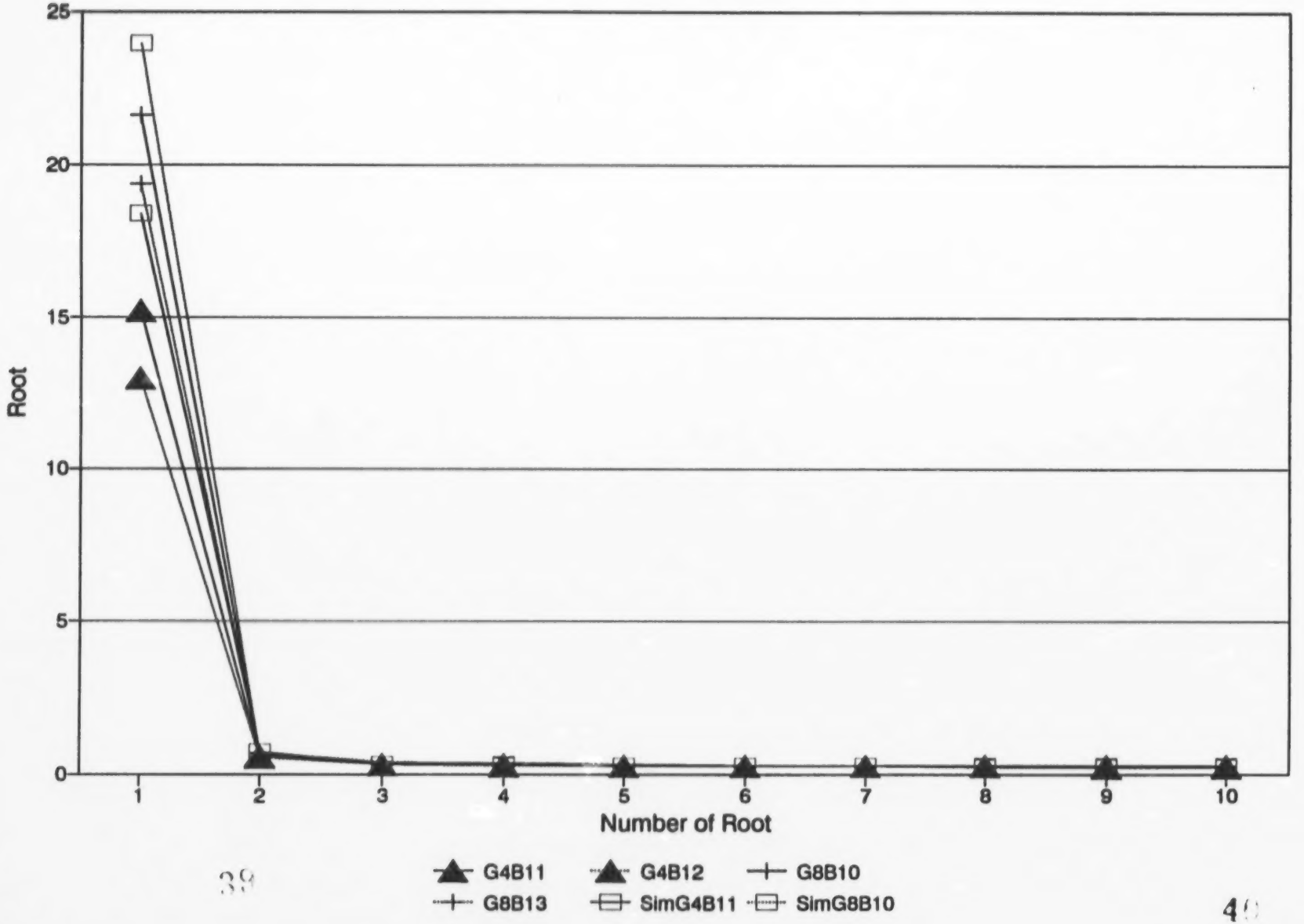


Figure 4a

Grade 4 Book 11 2nd thru 10th Roots

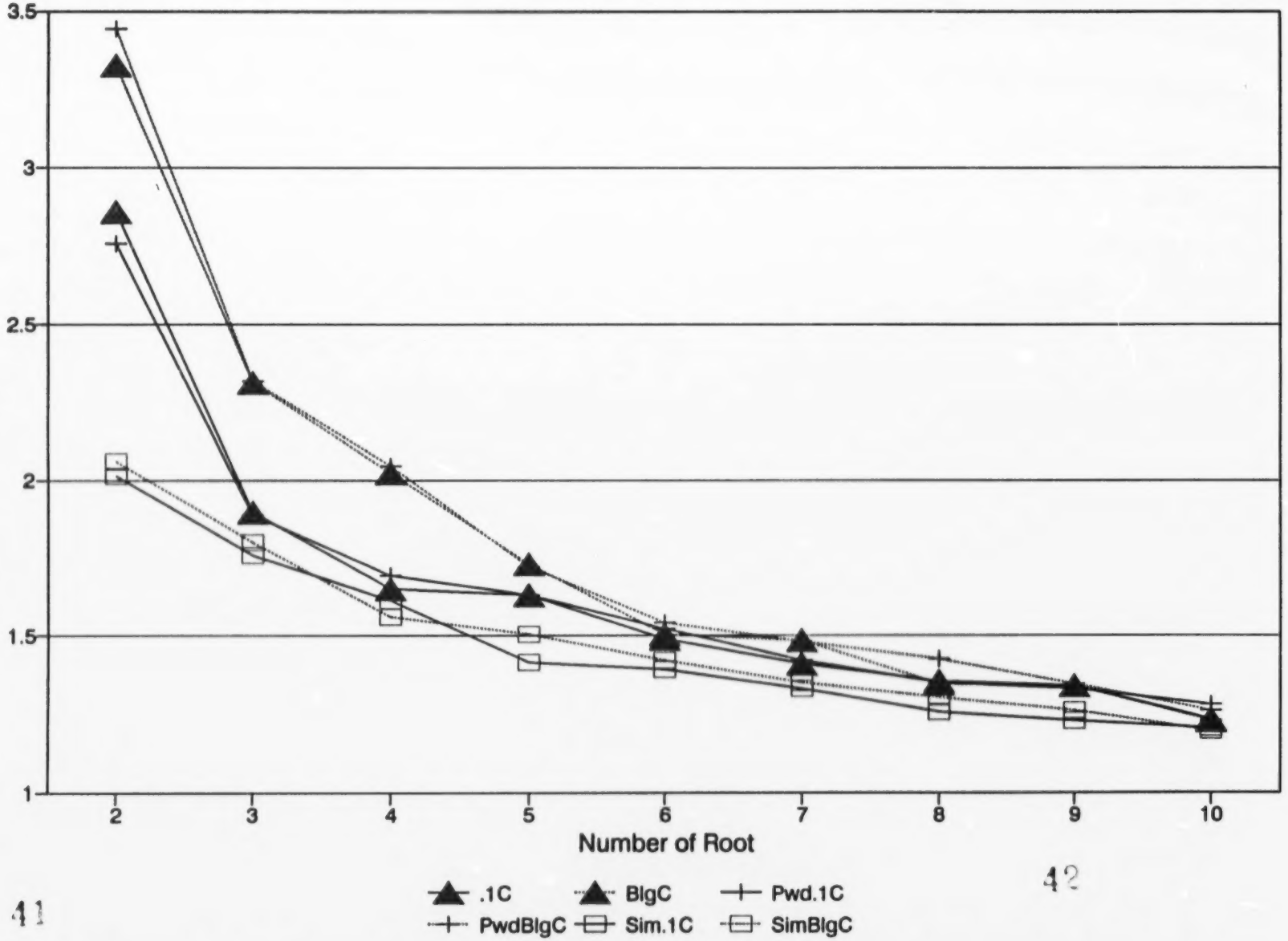


Figure 4b

Grade 4 Book 12 2nd thru 10th Roots

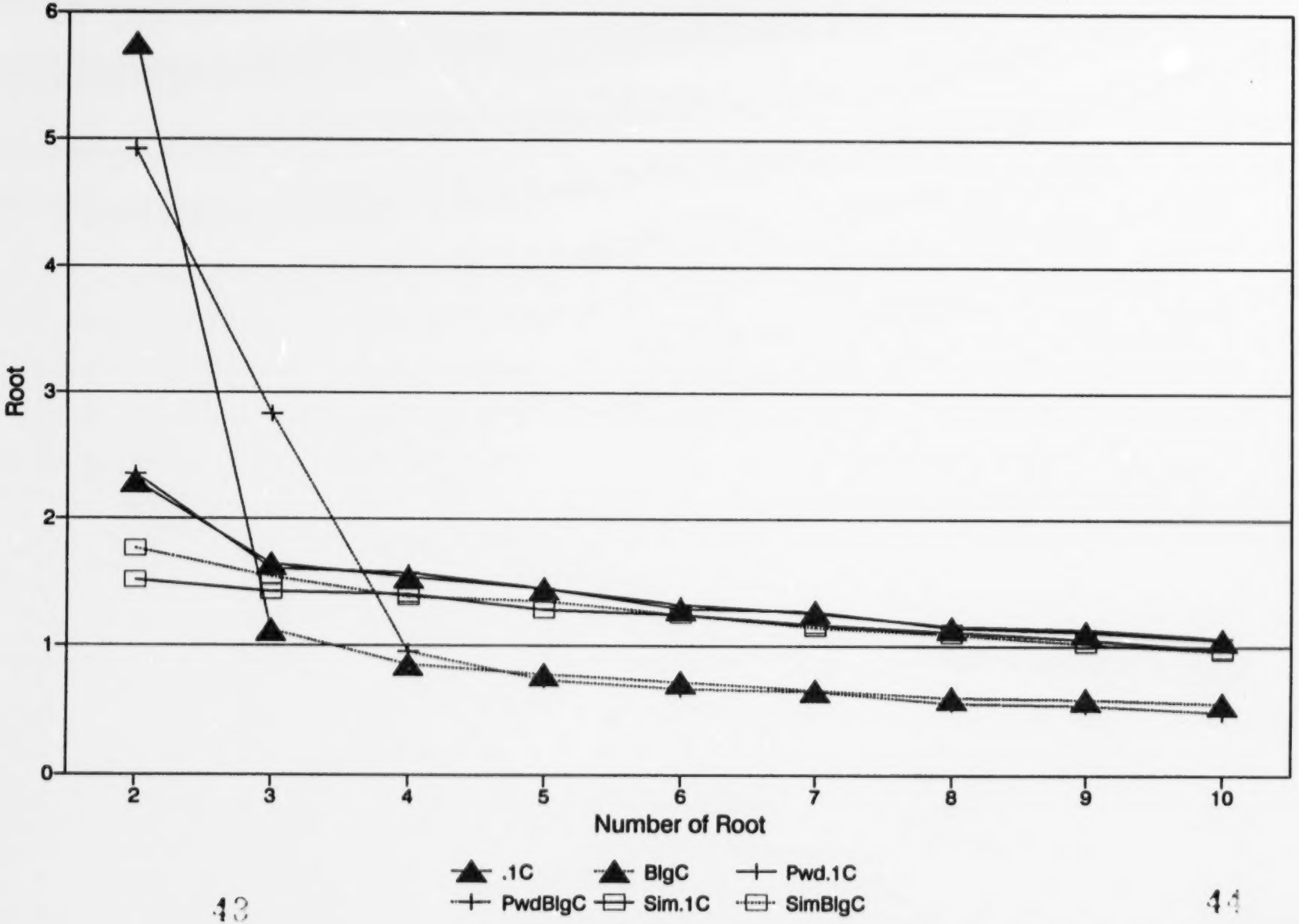


Figure 4c

Grade 8 Book 10 2nd thru 10th Roots

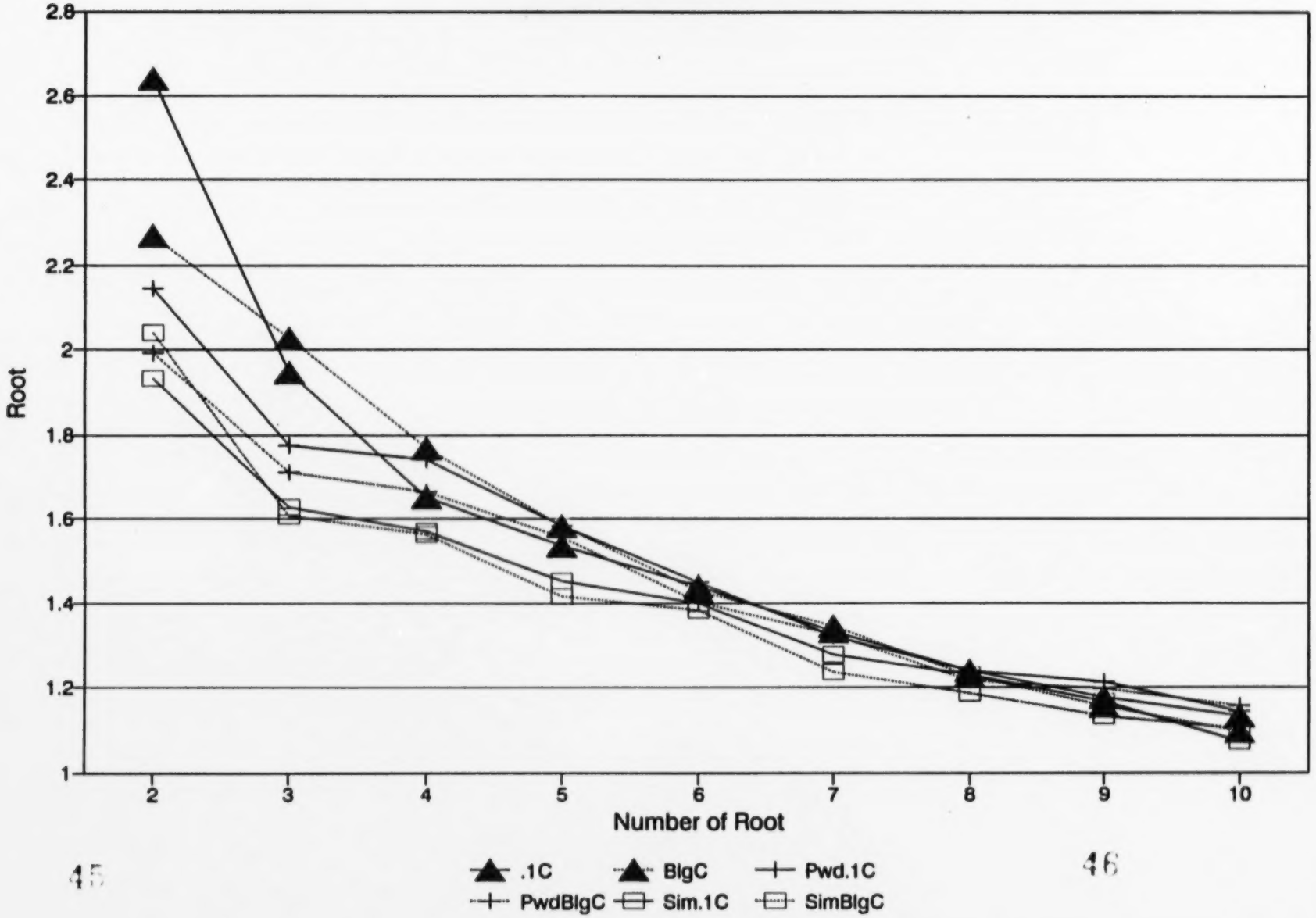


Figure 4d

Grade 8 Book 13 2nd thru 10th Roots

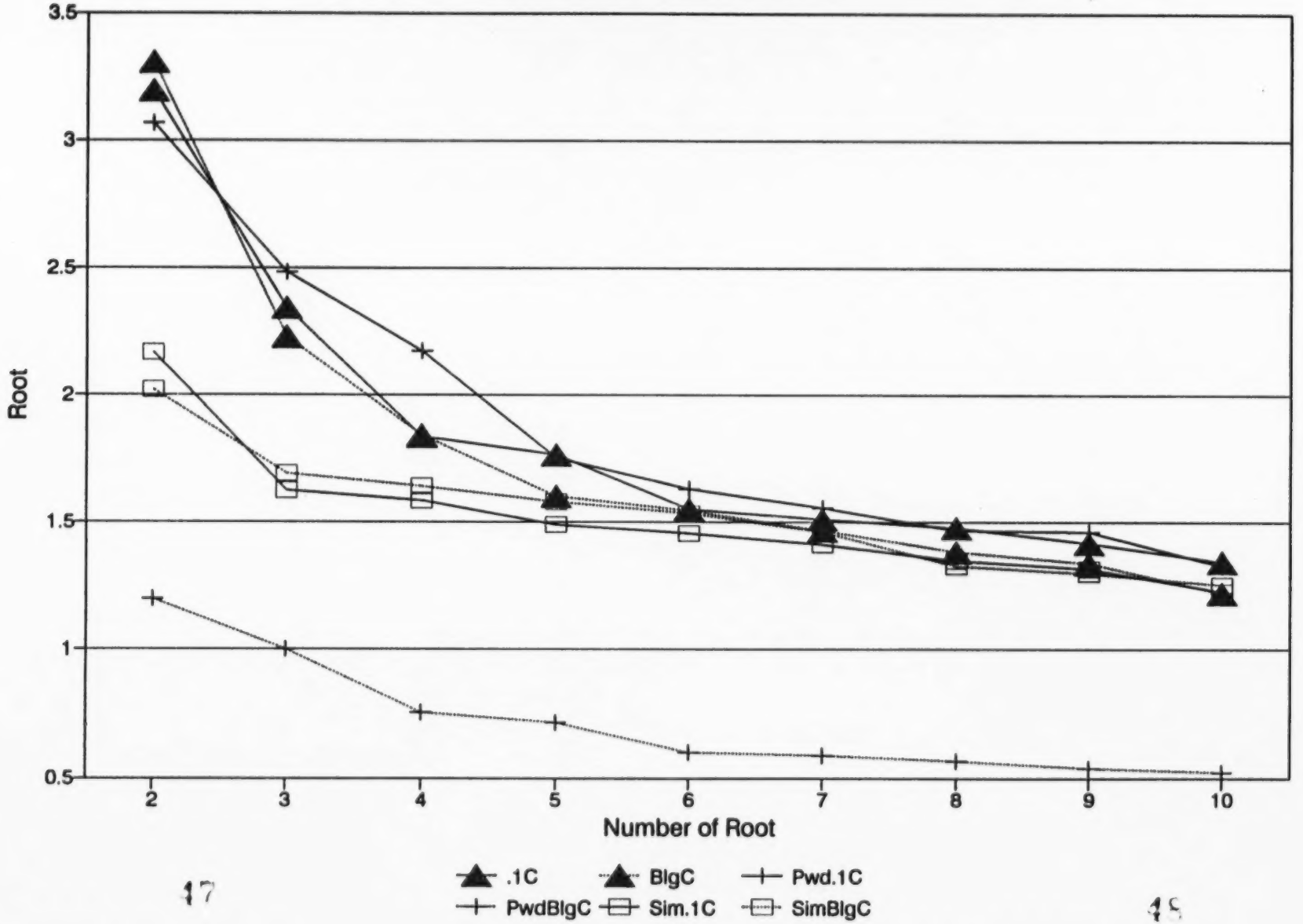
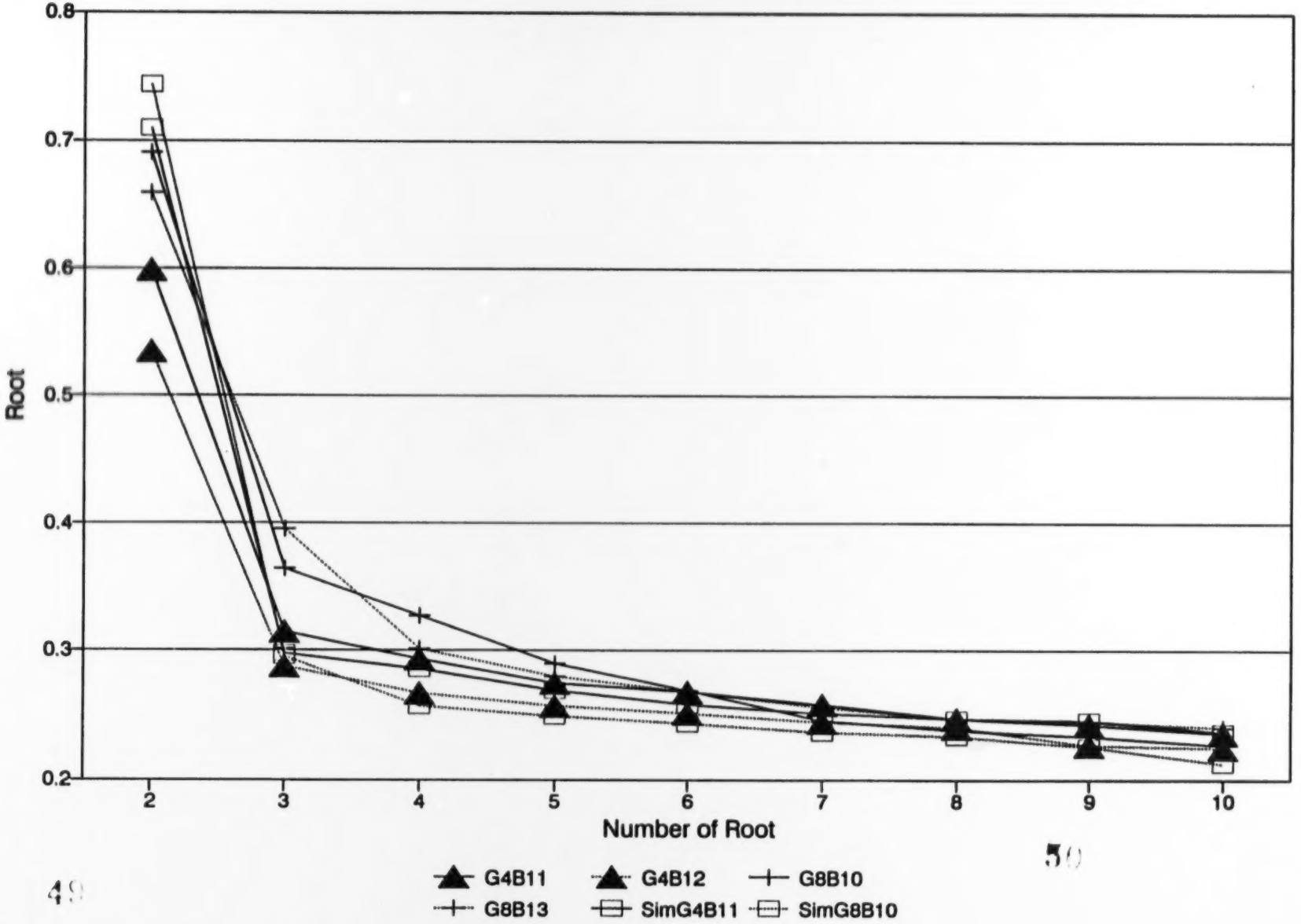


Figure 4e

Roots Two Thru Ten of SSCP/N



Additional Information

Further information about the measurement properties of the NAEP mathematics booklets that were analyzed is presented in Figures 5a through 5c and 6.

The mean tetrachoric coefficient for each of the matrix along with information about the variability of those coefficients is presented for each booklet in Figures 5a through 5c. The means are plotted together with a vertical line indicating the distance from the mean to plus and minus one standard deviation.

Figure 6 shows the sizes of the Kuder-Richardson reliability estimates as calculated by TESTFACT, for each of the datasets that was analyzed. As would be expected their sizes are highly related to the numbers of items (50 and 42 for the grade 4 booklets, 54 and 58 for the grade 8 booklets).

Discussion

Comparisons with simulated data show better fit when the data are known to be one-dimensional than occurs in the case of the real NAEP data. This result, which is hardly surprising, seems to be consistent over techniques (the NOHARM residual analysis and the TESTFACT procedures) and also over different datasets. There is, however, evidence of a strong first dimension fitting the examinee-item interaction data. Overall, the authors believe that the data are sufficiently unidimensional to support the use of a composite scale for describing the NAEP mathematics data, but that there is evidence that two dimensions would better fit the data than one.

Additional analyses are planned with these data. Current plans are to analyze some grade twelve data, and to try fitting some confirmatory factor analytic models to the data. Similar analyses will be performed on the science data from the main NAEP 1990 assessment.

Figure 5a

Grade 4 Book 11 Mean Tet., +/- S.D.

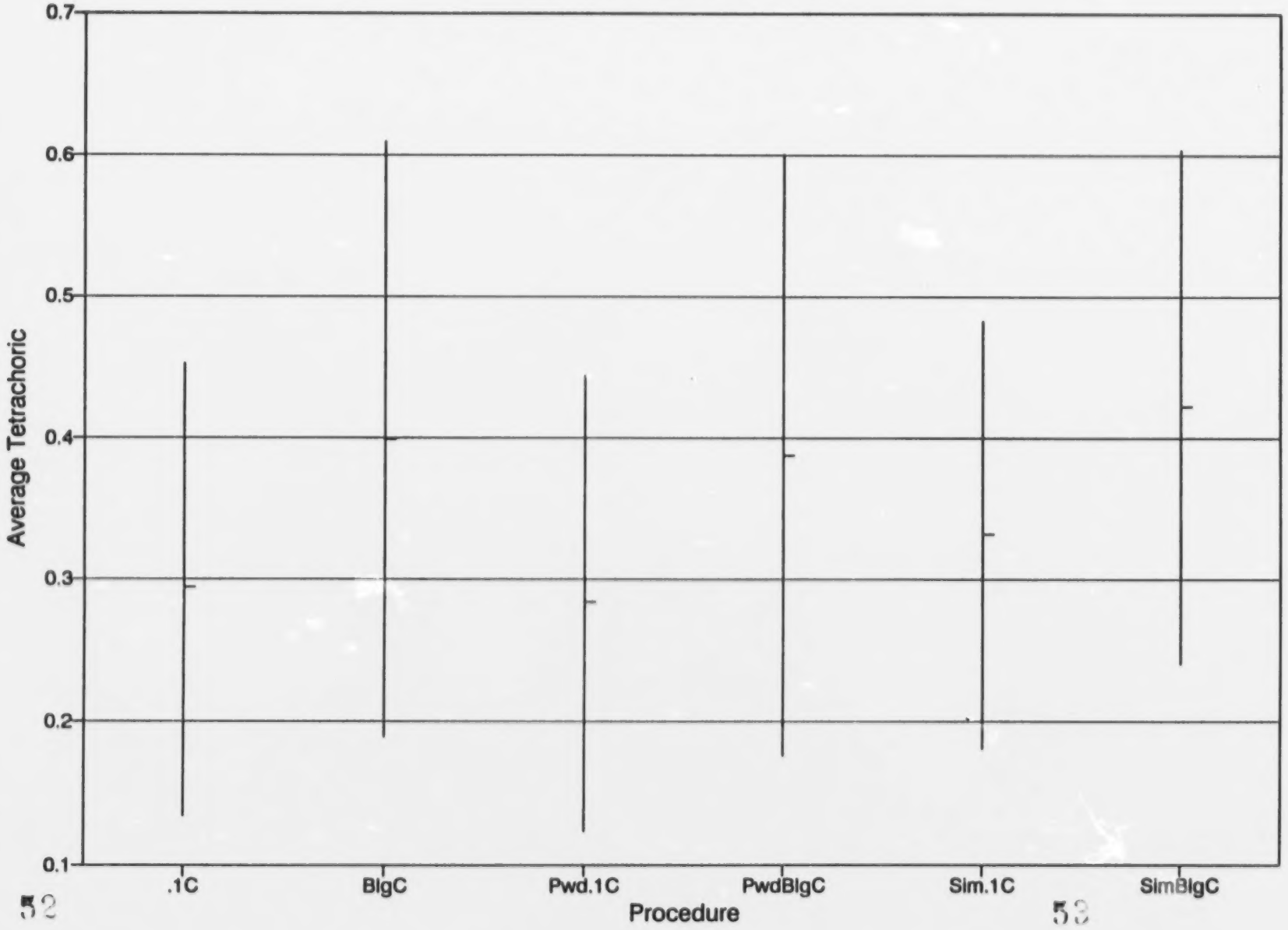


Figure 5b

Grade 4 Book 12 Mean Tet., +/- S.D.

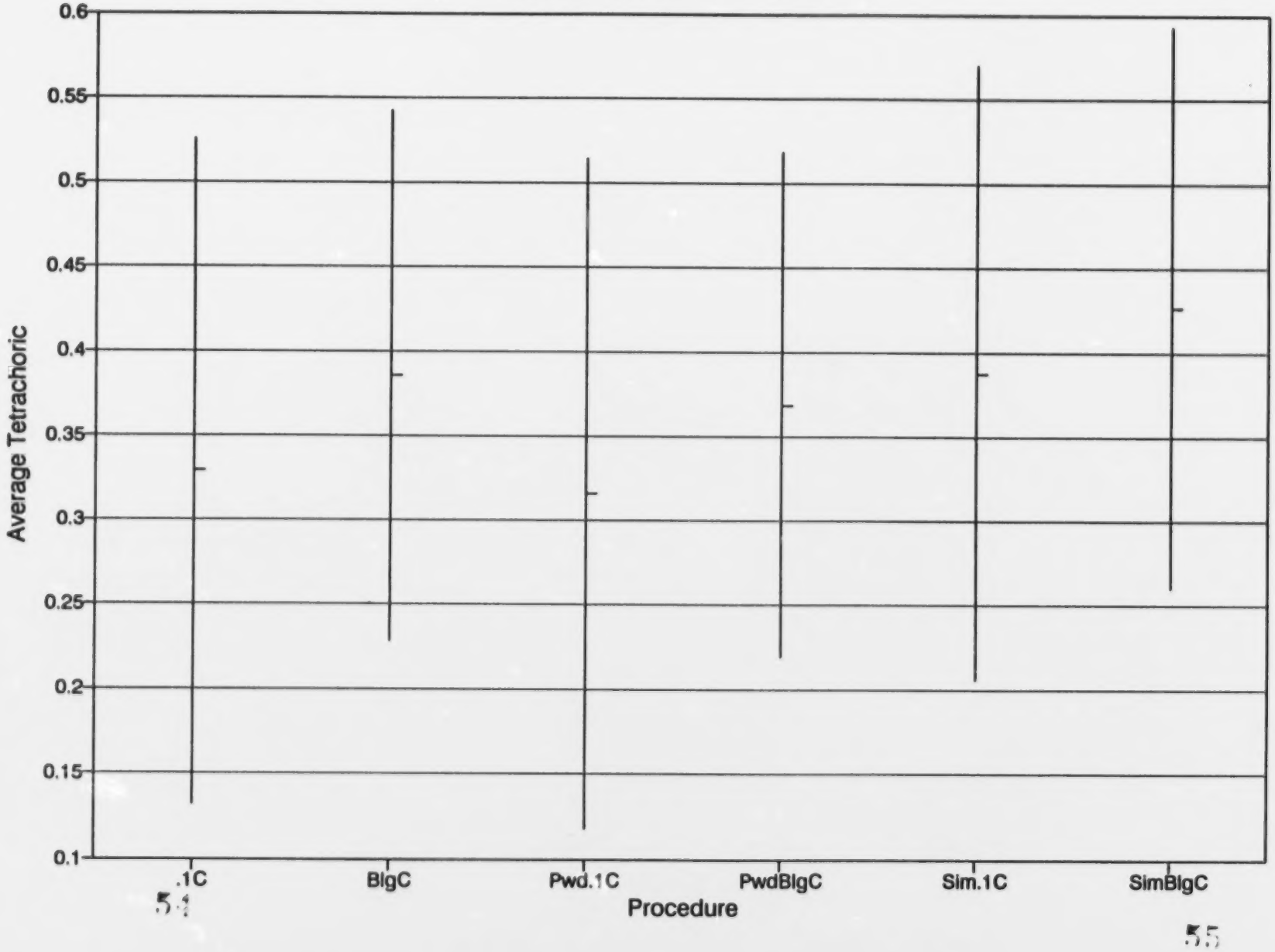


Figure 5c

Grade 8 Book 10 Mean Tet., +/- S.D.

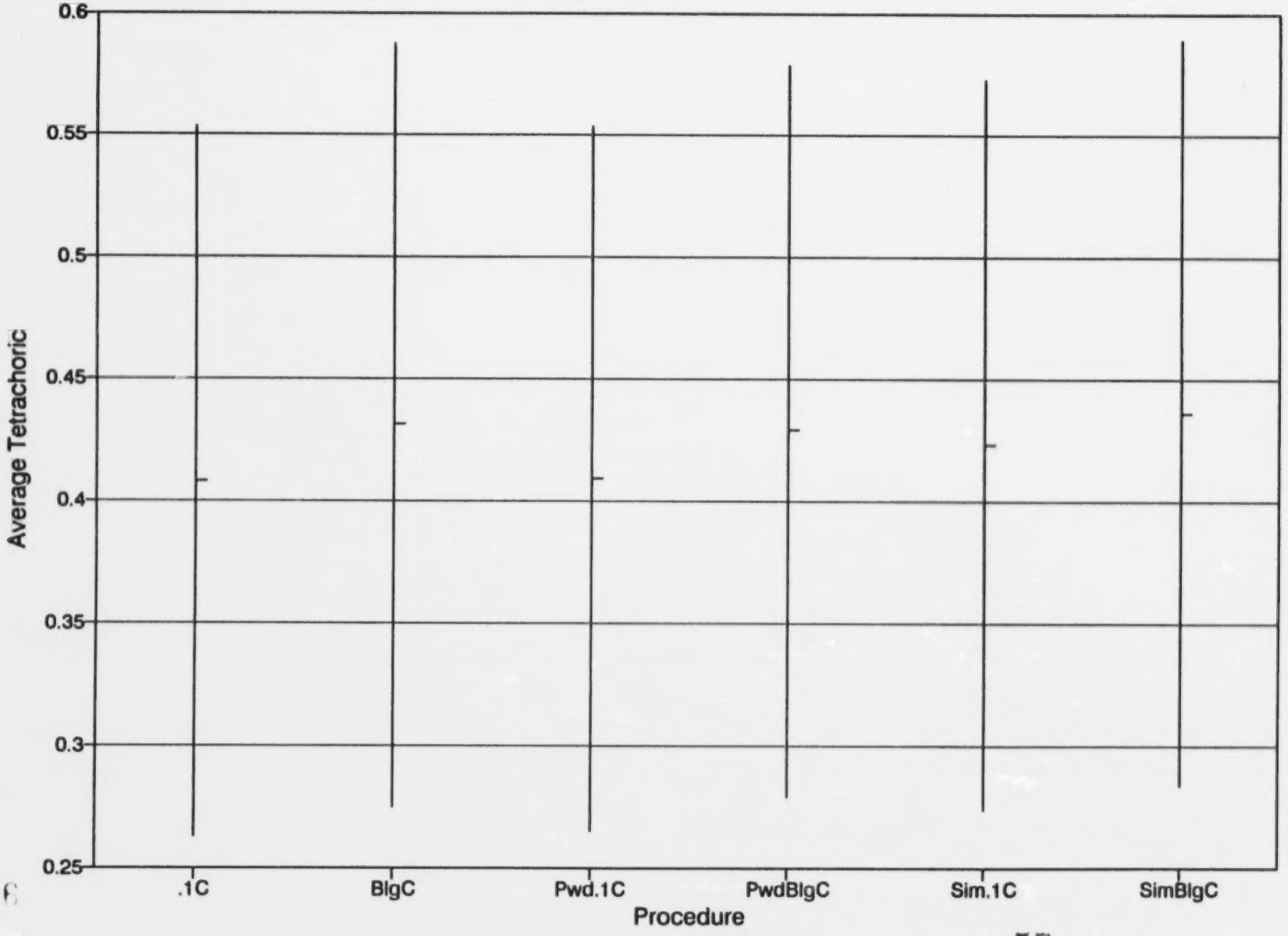


Figure 5d

Grade 8 Book 13 Mean Tet., +/- S.D.

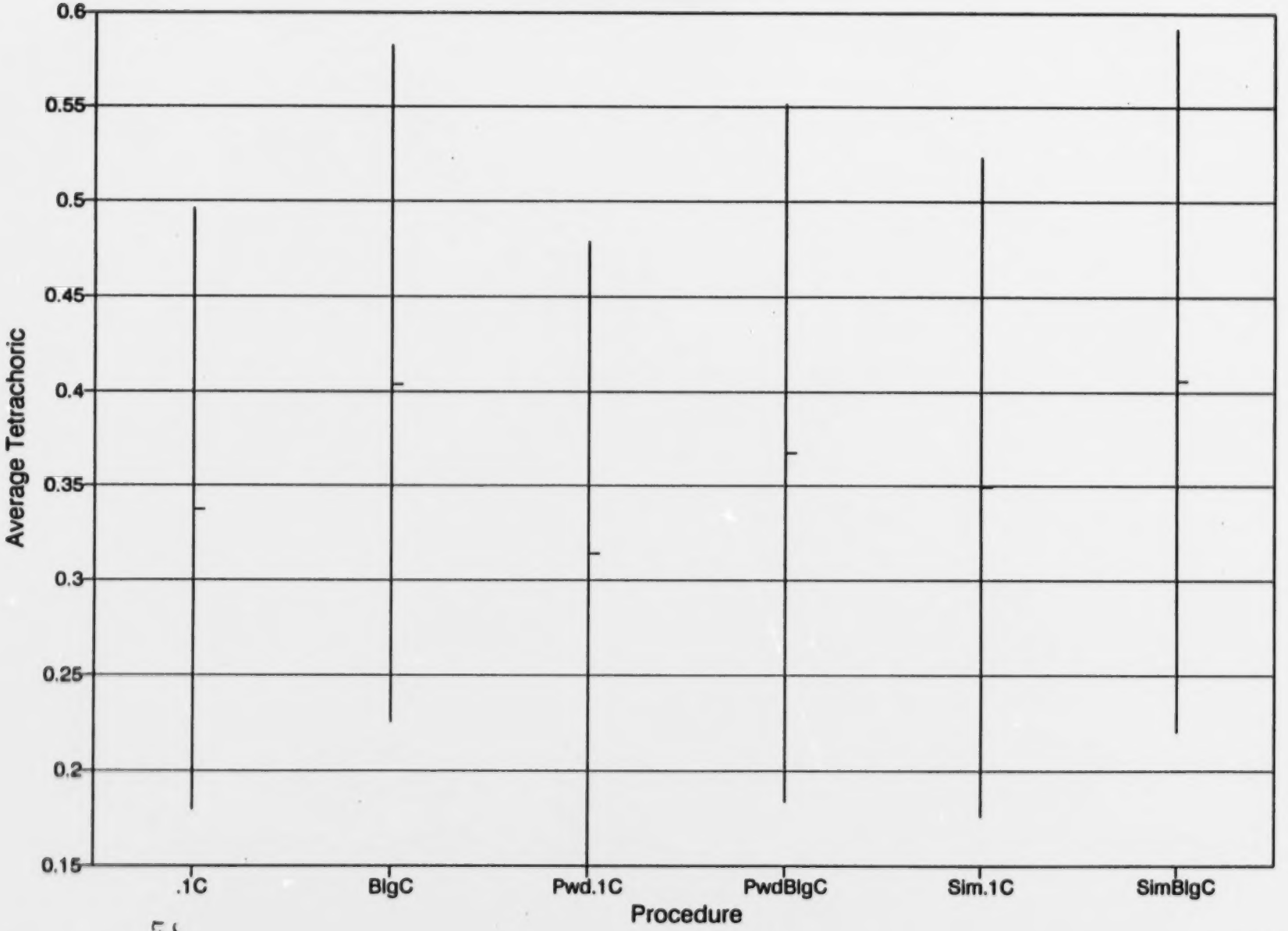
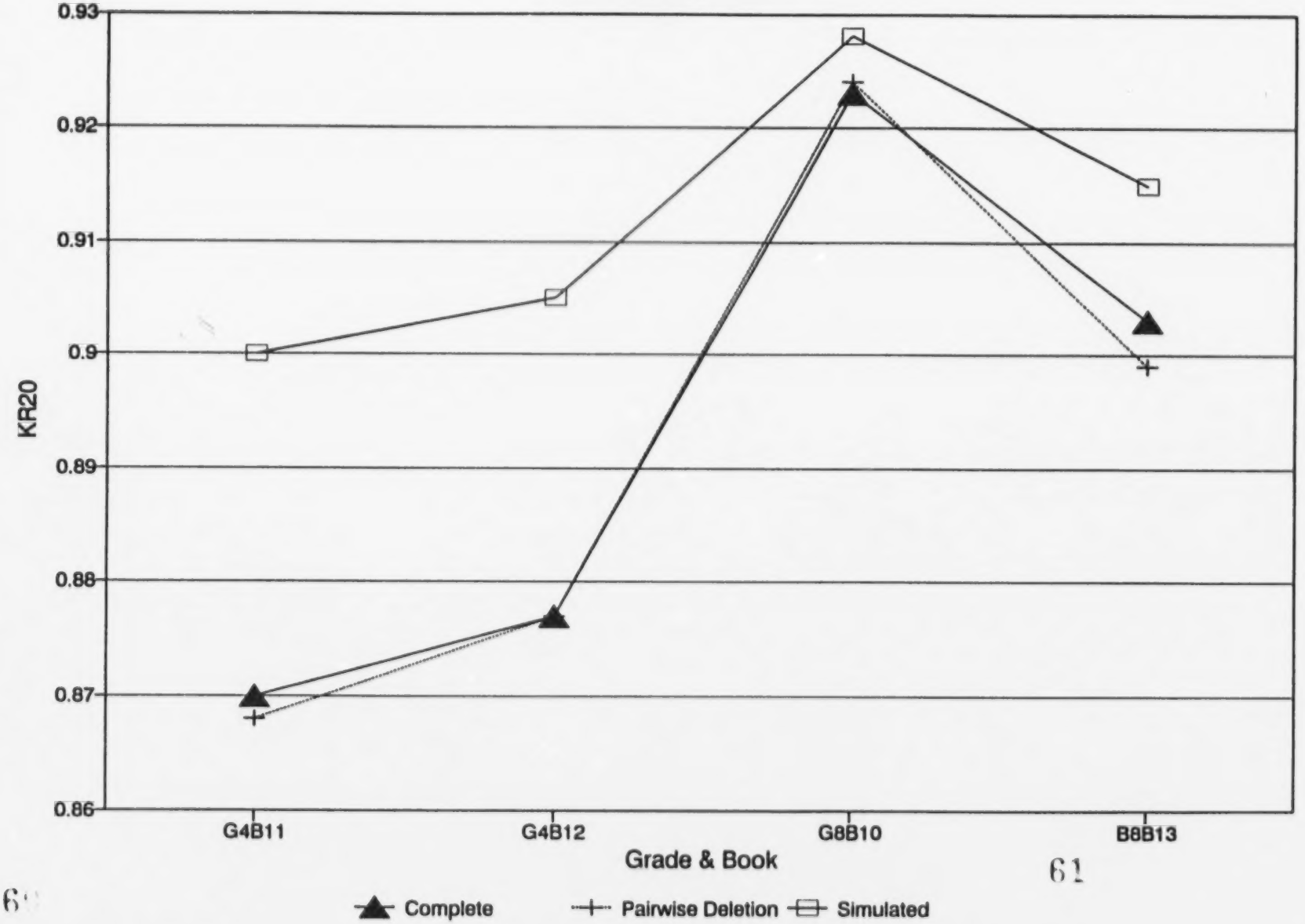


Figure 6

KR20 Reliability Coefficients



REFERENCES

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317-332.

Beaton, A. E., Johnson, E. G., & Ferris, J. J. (1987). The assignment of exercises to students. In A. E. Beaton, *Implementing the new design: The NAEP 1983-84 technical report* (pp. 97-118). Princeton, NJ: Educational Testing Service.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*, 261-280.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35*, 283-319.

Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika, 10*, 1-19.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40*, 5-32.

Divgi, D. P. (1979). Calculation of the tetrachoric correlation coefficient. *Psychometrika, 44*, 169-142.

Fraser, C. (1988). *NOHARM A computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory.* Armidale, Australia: The University of New England, Center for Behavioral Studies.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19*, 49-78.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139-164.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *Annals of Statistics, 134*, 1523-1543.

Johnson, E. G., & Allen, N. A. (1992). *The NAEP 1990 Technical Report*. Princeton, NJ: Educational Testing Service.

Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 26*, 457-477.

McDonald, R. P. (1962). A general approach to nonlinear factor analysis. *Psychometrika, 27*, 397-415.

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric Monographs*, No. 15.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100-117.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6*, 379-396. (a)

McDonald, R. P. (1982). Unidimensional and multidimensional models for item response theory. In D. J. Weiss (Ed.), *Proceedings of the 1982 Computerized Adaptive Testing Conference* (pp. 127-148). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program. (b)

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale, NJ: Lawrence Erlbaum.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11*, 3-31.

Mislevy, R. J., & Bock, R. D. (1982). *BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville, IN: Scientific Software.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43*, 551-560.

Muthén, B. (1991). *Issues in using NAEP mathematics items to study achievement dimensionality, within-grade differences, and across-grade growth*. Paper presented at the meeting of the Design and Analysis Committee of the National Assessment of Educational Progress, Watsonville, CA, December, 1991.

Rock, D. A. (1991). *Subscale dimensionality*. Paper presented at the meeting of the Design and Analysis Committee of the National Assessment of Educational Progress, Washington, DC, November, 1991.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49*, 425-436.

Rosenbaum, P. R. (1985). Comparing distributions of item responses for two groups. *British Journal of Mathematical and Statistical Psychology, 38*, 206-215.

Stout, W. (1983). A statistical test of unidimensionality of a parameter underlying bernoulli trails [sic] with applications. *Statistics & Probability Letters, 1*, 183-188.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293-325.

Wilson, D. T., Wood, R., & Gibbons, R. (1991). *TESTFACT: Test scoring, Item Statistics, and Item Factor Analysis 386/486 Version*. Mooresville, IN: Scientific Software.

Zwick, R. (1986). *Assessment of the dimensionality of NAEP year 15 reading data*. (ETS Research Report No. 86-4). Princeton: Educational Testing Service.

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement, 24*, 293-308.